White Paper Report

Report ID: 103287

Application Number: PR-50134-11

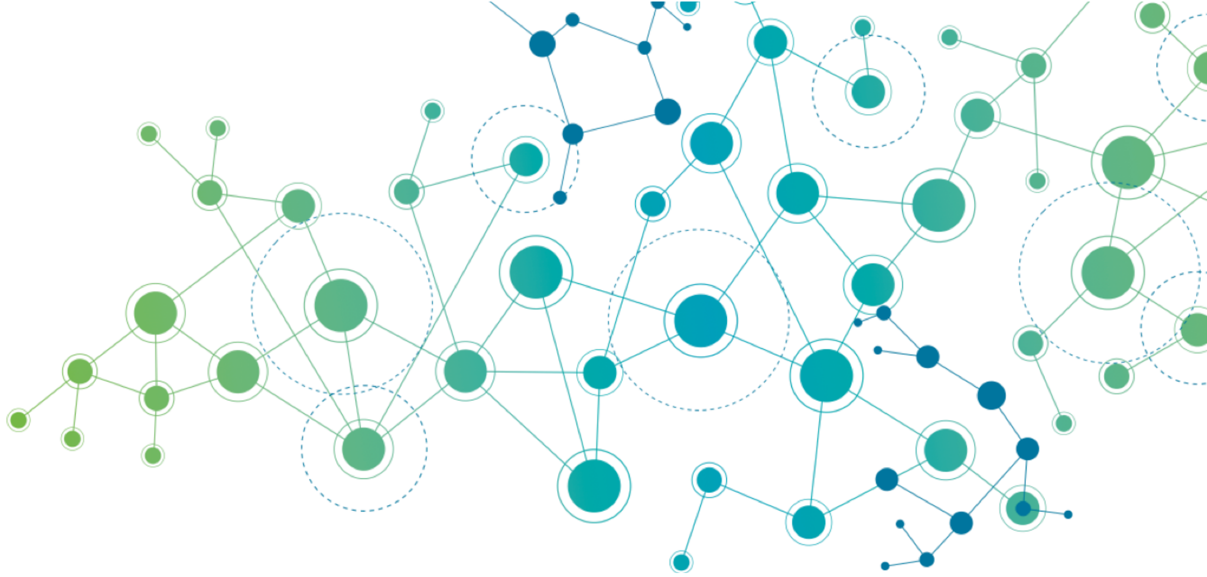
Project Director: Katherine Skinner (katherine.skinner@metaarchive.org)

Institution: Educopia Institute

Reporting Period: 8/1/2011-4/30/2014

Report Due: 7/31/2014

Date Submitted: 7/2/2014



Guidelines for Digital Newspaper Preservation Readiness

Katherine Skinner

Matt Schultz

2014 March 04

Version 1.0

Publication Notes:

Title: Guidelines for Digital Newspaper Preservation Readiness

Authors: Katherine Skinner, Matt Schultz

Publisher: Educopia Institute, 1230 Peachtree Suite, Suite 1900, Atlanta, GA 30309.

Table of Contents

Table of Contents

Introduction

Digital newspapers and preservation. Mention this topic to a roomful of curators, and in our experience, the conversation will move in a few predictable directions.

Most of the curators will share that they (or their institutions) manage some number of digital newspaper collections - mostly digitized, either in-house or by vendors. A few will share that they also manage born-digital content, either acquired from publishers directly or harvested via the Web.

Someone inevitably will point to the excellent set of standards in the field, particularly for digitization efforts.

And then…most curators will share stories of their own legacy collections, which conform only in small part or not at all to those standards. They will share tales of wildly variable silos of content that have been created under different management and/or varied grant-funded endeavors. These idiosyncratic and ad-hoc “collections” of digital news are inconsistent in their file types, structures, metadata, and storage locations. And their curators worry a lot about how to preserve these collections for future generations.

If you ask this hypothetical roomful of curators why they haven’t readied this content for preservation, most (if not all) will cite the same major barrier: resources. They have limited resources to expend on remediation activities, and the specifications and standards they might ideally deploy are simply too great an investment to consider.

The Educopia Institute and its affiliated digital preservation program, the MetaArchive Cooperative, have heard (and hosted) these kinds of conversations for several years. In 2009 and 2010, we ran surveys to verify and document the needs we were hearing. And then, with generous funding from the National Endowment for the Humanities and in partnership with a range of experts from Chronopolis and the libraries of University of North Texas, Penn State, Virginia Tech, University of Utah, Georgia Tech, and Boston College, we started to explore ways to lower the barrier to entry for managing digital newspapers for the long term.

These *Guidelines* are an effort to distill preservation-readiness steps into incremental processes that an institution of almost any size or type can deploy to begin maturing its digital newspaper content management practices.

We wish to thank the project team and its Advisory Board, including Mary Molinaro, Sue Kellerman, Liz Bishoff, and Frederick Zarndt, for the editorial feedback they provided at multiple stages of our development of these *Guidelines*. We also wish to thank those that weighed in during an open review period we hosted for these documents in 2013 - we were glad to learn from your experiences, and we incorporated your suggestions into this final version. Finally, we wish to thank Nick Krabbenhoeft, who helped us immensely with editing and refining these documents.

Any oversights and omissions herein are entirely our own, and we'll point out the most grave: the *Guidelines* only deal with digital newspapers at this point, not broadcast or other forms of digital news. AV preservation brings in a host of additional factors and considerations. We hope to later expand the *Guidelines* (or to encourage someone else to do so…) to include the broader spectrum of "digital news," but we are beginning with what we know well and can document most thoroughly: digital newspapers. The *Guidelines* are herein at Version 1.0 and we hope to contribute to future versions that can track the community's progress with practices and technologies.

We hope to hear from you!

Katherine Skinner and Matt Schultz

About the Guidelines

The *Guidelines for Digital Newspaper Preservation Readiness* address a specific set of preservation challenges faced by libraries, archives, historical societies, and other organizations that curate substantial collections of digital newspaper content. The digital newspaper collections managed by these memory organizations often have been established over several decades of digitization and born-digital acquisition efforts. As such, they tend to encompass a wide range of file types, structures, and metadata schemas. With limited staffing, time, and infrastructure, how can institutions prepare such diverse collections for preservation?

Consider just a few example scenarios:

- A state historical society has digitized hundreds of newspaper pages through its participation in programs such as the United States Newspaper Program (USNP) and the National Digital Newspaper Program (NDNP). With significant federal funding, these public domain newspapers have been catalogued, digitized, transcribed, and modeled according to evolving best practices in metadata, imaging, optical character recognition (OCR), and other standards and technologies. Due to these changing standards and technologies, the historical society now must maintain legacy collections that were digitized according to different "best practices". As such, it must determine how best to streamline these diverse collections into a set of content that can be managed over time.
- An academic research library has acquired the digital back content of a local commercial news publisher, including daily articles that were published on the Web in the mid-1990s using early versions of HTML. The library aims to serve this content out to various communities over time, which necessitates attention to copyright issues, server/operating system infrastructure changes, maintaining thousands of file linkages, and ongoing questions regarding how different browsers will render this information as such technologies and their underlying standards evolve.
- A state library has started to collect a local publisher's print-ready files and Web content - including social media feeds. The library must work out agreements and processes for how best to acquire authoritative and complete content from the publisher on a routine basis. It also must make decisions regarding migration and normalization to ensure the longevity of the content and to facilitate its integration with the state library's extensive digitized newspaper holdings.

Digital newspaper collections are a key historical record of human activities. Given the preservation challenges posed by this valuable and unique set of scholarly assets, curators are asking the question "How can we effectively and efficiently prepare our digitized and born-digital newspaper collections for preservation?"

These *Guidelines* are intended to inform curators and collection managers at libraries, archives, historical societies, and other such memory organizations about various practical readiness activities that they can take. They provide links to technical resources that curators can either implement themselves or work with their technical staff to implement (for more, see *How to Use the Guidelines* below).

What is "Digital Preservation"?

It is important to understand at the outset what we mean when we use the term "digital preservation." Digital preservation is widely understood as the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary."[1] An operative term in this definition is "managed activities." Institutions seeking to preserve digital materials must understand that preservation requires planning, care, and coordination over time. The definition does not seek to quantify the purpose, scope, or duration of preservation - only marking it "for as long as necessary." Which is to say, institutions usually do not have a mandate to preserve everything forever. Digital preservation is instead an ongoing process that can be undertaken pragmatically and incrementally.

Understanding Standards for Digital Newspapers

Digital preservation standards and practices grow annually in number and complexity. As a result, it is hard to know where to start, or how to define workflows that will last a reasonable length of time.

An institution that seeks to preserve its digital newspapers may turn to various authoritative sources to get its bearings. It might turn to the Library of Congress[2] or the newspaper sections and working groups of various library and archival professional associations such as the *Center for Research Libraries Global Resources Network (CRL GRN)*,[3] the *American Library Association (ALA)*,[4] the *Society of American Archivists (SAA)*,[5] or the *International Federation of Library Associations (IFLA)*.[6] It might also turn to various professional listservs such as *newslib*,[7] *digi-pres*,[8] *code4lib*,[9] or *digital-curation*.[10]

Along the way, an institution almost certainly will gain familiarity with the standards known as

[1] Digital Preservation Coalition, "Introduction – Definitions and Concepts," available at: http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts.

[2] Library of Congress, "Digital Preservation," available at: http://www.digitalpreservation.gov/.

[3] Center for Research Libraries Global Resources Network, "Global Resources Program," available at: http://www.crl.edu/grn.

[4] Association for Library Collections & Technical Services, "Newspaper IG," available at: http://www.ala.org/alcts/mgrps/ig/ats-dgnews.

[5] Society of American Archivists, "Homepage," available at: http://www2.archivists.org/.

[6] International Federation of Library Associations, "Newspapers Section," available at: http://www.ifla.org/newspapers.

[7] Newslib, "Homepage," available at: http://www.ibiblio.org/slanews/NewsLib/newsliblyris.html.

[8] American Library Association, "digipres- Digital Preservation," available at: http://lists.ala.org/sympa/info/digipres.

[9] Code4Lib, "Homepage," available at: http://www.lsoft.com/scripts/wl.exe?SL1=CODE4LIB&H=LISTSERV.ND.EDU.

[10] Google Groups, "Digital Curation-Google Groups," available at: https://groups.google.com/forum/#!forum/digital-curation.

the *Reference Model for an Open Archival Information System (OAIS)*,[11] and *ISO:16363 Audit and certification of trustworthy digital repositories*.[12] Both of these standards have been instrumental in formulating the general concepts and terminology necessary to implement a digital archive. They also help to outline the organizational and technical aspects that auditors and stakeholders should be able to evaluate. These standards aim less to suggest particular implementations than to set forth the full range of requirements needed to accomplish preservation in a responsible fashion.

An institution likely will also encounter the *National Digital Newspaper Program (NDNP) Technical Guidelines*.[13] Released first in 2007 and updated for each phase of NDNP, these specifications address scanning resolutions and establish standard, high-quality file formats for digitization (e.g., TIFF 6.0). They also provide quality requirements for uniform metadata (e.g., CONSER-derived), encoding levels (METS-ALTO), and derivative file formats (e.g., JPEG2000 and PDF w/Hidden Text). Each of these technical requirements is in keeping with current, accepted high standards for image-based archival-quality digitization and prepares the collections for long-term preservation.

An institution will also grapple with various recommendations regarding preservation metadata standards and schemas. In particular, two standards - the *Metadata Encoding Transmission Standard (METS)*[14] and *Preservation Metadata: Implementation Strategies (PREMIS)*[15] - have been designed as robust strategies for encapsulating the widest possible range of preservation-oriented information about digital objects and collections. The goal of these standards is to help institutions provide better lifecycle management for digital objects.

Each of these standards documents comprehensive strategies for accomplishing some part of the complex task of preserving digital content. However, these comprehensive standards can seem formidable, even to experienced preservationists. Upon gaining familiarity with the standards literature, an institution might worry that it would need to completely re-think or reverse its practices to begin preserving its content.

If an institution can engage in an incremental process that allows it to begin preserving content now, while slowly and steadily building toward an optimal level of preservation

[11] Consultative Committee for Space Data Systems, *CCSDS 650.0-M-2: Reference Model for an Open Archival Information System (OAIS): Magenta Book,* June 2012, available at: http://public.ccsds.org/publications/archive/650x0m2.pdf.

[12] CCSDS, “ISO 16363:2012 Audit and certification of trustworthy digital repositories – Magenta Book,” available at: http://public.ccsds.org/publications/archive/652x0m1.pdf.

[13] Library of Congress, “The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants,” August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

[14] Library of Congress, “Metadata Encoding Transmission Standard (METS),” available at: http://www.loc.gov/standards/mets/.

[15] Library of Congress, “Preservation Metadata: Implementation Strategies (PREMIS),” available at: http://www.loc.gov/standards/premis/.

readiness, it will be more likely to begin participating in preservation activities. Once institutions begin preserving content, they also will begin building the requisite expertise and knowledge in this area to prepare new collections and normalize legacy collections according to optimal standards.

A Preservation Spectrum - Essential to Optimal

These *Guidelines for Digital Newspaper Preservation Readiness* aim to explicitly differentiate between the *essential* and *optimal* in preservation readiness activities and document the incremental steps that institutions may take to move from the *essential* to the *optimal* level of preservation readiness for their digital newspapers.

By *essential* we mean:

- practices that are reasonable to accomplish given a limited set of resources and expertise; and
- practices that are non-negotiable because to neglect them would be to ignore preservation.

By *optimal* we mean:

- practices that are reasonable to expect given an ample set of resources and expertise; and
- practices that can ensure the most reliable long-term preservation.

The Need for a Preservation Spectrum

Digital newspapers span a diversity of forms. There are newspapers that consist of page images in digital microfilm format, newspapers that have been digitally scanned from analog microfilm and from print (at various image resolutions), encoded text derived from these scanned images (optical character recognition or OCR), and of course born-digital newspapers - often e-prints and web-related text, image, and multimedia files.

Because digital newspaper files are created under such a wide range of circumstances, from grant-funded projects to ad hoc scanning initiatives, they also tend to be stored on a variety of media. In a given library or archive, digital newspaper files might be found on CDs, portable hard-drives, tape back-up systems, and various flavors of disk arrays.

Finally, a range of institutional types curate digital newspaper collections. These run the gamut from public libraries to historical societies, museums to academic libraries, and state libraries to vendor groups. Each of these memory stewards has slightly different contexts within which it acquires, creates, and manages digital newspaper content, and depending on its wherewithal and good fortune (or the lack thereof), each has more or fewer resources to put behind preserving its assets.

All of this underscores the drivers for producing the *Guidelines*: namely that all institutions can do something to prepare their collections for long-term use, and that there can be no one-size-fits-all approach to preserving digital newspapers. Institutions need to be able to tackle the challenges involved in preserving digital newspapers in modular increments. Though they need to be able to understand the entire series of "managed activities" as inter-related stepping stones, they also need to be empowered to produce staged implementations based on their current and future capacities.

Reference: Digital Preservation Standards

The following standards are relevant to preserving digital newspapers and for guiding digital preservation practices more generally:

The *National Digital Newspaper Program (NDNP) Technical Guidelines*[16] describe the specific technical requirements for inclusion of digital content in Chronicling America (NEH/Library of Congress). They aim to support the "best practices" of today's understanding of digital preservation needs for digitized newspapers.

The *Reference Model for an Open Archival Information System (OAIS)* was developed under the auspices of the Consultative Committee for Space Data Systems (CCSDS) and first approved as an ISO standard in 2003 (ISO 14721:2003). It is now superseded by ISO 14721:2012.[17] The Reference Model addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, and dissemination.

ISO 16363:2012 Audit and certification of trustworthy digital repositories[18] defines a recommended practice for assessing the trustworthiness of digital repositories. It builds off of *Trustworthy Repositories Audit & Certification: Criteria & Checklist* Version 1.0, 2007.[19]

The Metadata Encoding & Transmission Standard (METS)[20] is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.

The *Preservation Metadata Implementation Strategies (PREMIS)*[21] data dictionary is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability.

[16] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants," August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

[17] Consultative Committee for Space Data Systems, *CCSDS 650.0-M-2: Reference Model for an Open Archival Information System (OAIS): Magenta Book,* June 2012, available at: http://public.ccsds.org/publications/archive/650x0m2.pdf.

[18] CCSDS, "ISO 16363:2012 Audit and certification of trustworthy digital repositories – Magenta Book," available at: http://public.ccsds.org/publications/archive/652x0m1.pdf.

[19] Center for Research Libraries, "Trustworthy Repositories Audit & Certification: Criteria & Checklist," February 2007, available at: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

[20] Library of Congress, "Metadata Encoding Transmission Standard (METS)," available at: http://www.loc.gov/standards/mets/.

[21] Library of Congress, "Preservation Metadata: Implementation Strategies (PREMIS)," available at: http://www.loc.gov/standards/premis/.

Trends Towards Incremental Approaches

The *Guidelines* are not alone in seeking to address the need for preservation spectrums. The *National Digital Stewardship Alliance (NDSA)* is a U.S.-based, member-driven association that seeks to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations.[22] NDSA coordinates working groups that include curators from both public and private organizations who seek to advance digital preservation for both their own organizations and for the community at large.

Under the auspices of the NDSA, the Innovation Working Group has established a spectrum-based approach termed the *Levels of Preservation*.[23]

In 2014, the *Levels of Preservation* are at version one, and are described as:

> *[A]* tiered set of recommendations for how organizations should begin to build or enhance their digital preservation activities. A work in progress by the NDSA, it is intended to be a relatively easy-to-use set of guidelines useful not only for those just beginning to think about preserving their digital assets, but also for institutions planning the next steps in enhancing their existing digital preservation systems and workflows….The guidelines are organized into five functional areas that are at the heart of digital preservation systems: storage and geographic location, file fixity and data integrity, information security, metadata, and file formats.

There is a growing recognition in the field that building institutional capacity in digital preservation and curation may be best achieved through manageable and incremental approaches.

The *Guidelines for Digital Newspaper Preservation Readiness* do not seek to apply the NDSA Levels of Preservation to newspaper content directly, but these "Levels" are consistent with our "spectrum" approach and are an important resource for institutions embarking upon (or refining existing) preservation pathways.

How to Use the Guidelines

Two points must be addressed up front regarding the purpose of the *Guidelines*.

First, the authors understand that typically, this type of documentation begins with selection, acquisition, and/or creation. The *Guidelines* do not, because they are intended for curators who *already have collections that need preservation attention*. We want to encourage institutions to dive in and begin actively pursuing preservation today with existing, at-risk content. We do not, however, ignore selection, acquisition, and/or creation as critical components of lifecycle management. We address them at the end, rather than the beginning, of the *Guidelines.* There, we provide brief recommendations for concrete steps that newspaper stewards may take to integrate and incorporate preservation readiness as they

About the Guidelines

[22] Library of Congress, "National Digital Stewardship Alliance Homepage," available at: http://www.digitalpreservation.gov/ndsa/.

[23] Library of Congress, "NDSA Levels of Preservation," available at: http://www.digitalpreservation.gov/ndsa/activities/levels.html.

Reference: NDSA Levels of Preservation

The goal of the *NDSA Levels of Preservation* is to provide a basic tool for helping organizations manage and mitigate digital preservation risks.

The Levels make use of a matrix that plots standard digital preservation activities such as storage, file fixity, information security, metadata, and file formats within each of four Levels. The Levels walk an institution forward through 1) Protecting Your Data; 2) Knowing Your Data; 3) Monitoring Your Data; and 4) Repairing Your Data.

An example of how the Levels work for metadata is below:

	Level One (Protect Your Data)	Level Two (Know Your Data)	Level Three (Monitor Your Data)	Level Four (Repair Your Data)
Metadata	-Inventory of content /storage locations -Offsite backup of inventory	-Store standard administrative and transformative metadata -Log events	-Store standard technical & descriptive metadata	-Store standard preservation metadata

Learn more at: http://www.digitalpreservation.gov/ndsa/activities/levels.html.

continue to create and acquire digital newspapers. We point outward to other, deeper resources regarding the field's current practices in each area - in particular, the mature digitization standards and specifications for newspapers (see *Section 7.1. Additional Considerations: Creation & Acquisition*).

We also highlight a key acquisition/creation pathway that is still emerging - that of born-digital newspaper collection capture and management. We describe some of the steps we believe memory organizations should be taking to ensure the survival of this content and point to future research that groups such the Center for Research Libraries, the National Digital Stewardship Alliance, and Educopia Institute are undertaking in this area.

Secondly, the *Guidelines* focus primarily on *preservation*, not *access*. The *Guidelines* intentionally separate these two functions, though its authors acknowledge the deep connections between them. What we preserve, we always should preserve so that it may be used someday by someone. With that emphasis established, the *Guidelines* aim first to break “preservation” down into a manageable set of modular preservation readiness activities. Given adequate resources, an institution could use these in a sequential fashion to produce a preservation program. In that way the *Guidelines* can be engaged as a roadmap to structure an institution’s digital newspaper curation activities from day one through to final packaging for long-term preservation (in OAIS terms, the creation of a Submission Information Package). A *Roadmap Checklist* is included with the *Guidelines* on page xi for just such approaches.

From top-to-bottom the *Guidelines* address the following preservation readiness activities:

- *Inventorying Digital Newspapers for Preservation*
- *Organizing Digital Newspapers for Preservation*
- *Format Management for Digital Newspapers*
- *Metadata Packaging for Digital Newspapers*
- *Checksum Management for Digital Newspapers*
- *Packaging Digital Newspapers for Preservation*

However, the *Guidelines* are also written with the full understanding that comprehensive and sweeping actions are often beyond the capacity of institutions that curate digital newspaper content. For that reason, each spotlighted preservation readiness activity is given its own section or module. Each module explains a core facet of digital preservation, unpacks its rationale, demonstrates how it can be applied to newspapers, and provides an overview of tools and methodologies.

In addition, because digital newspapers span both digitized and born-digital forms, the *Guidelines* attempt to call attention to how each of the preservation readiness activities may need to be tailored to address the unique needs of each form. In the foreground of each section is the broader set of "managed activities" that define digital preservation.

Where possible we have included case studies and examples to demonstrate how a real institution has engaged a particular readiness activity. These case studies celebrate successes, but also highlight challenges and share insights into decision-making around a preservation readiness activity given the institution's available (and often limited) resources.

Finally, and most importantly, each preservation readiness section includes a sub-section that situates the activity in the context of the suggested spectrum of *essential* and *optimal* practices. This sub-section will help institutions understand where they land on this spectrum, and advise how best to proceed with the recommended practices, tools, and methodologies. It is important for curators and collection managers to share some of the more technical elements that are discussed in the *Guidelines* with their technical staff members and consultants, who may help to expand upon some of the tools and implementations that are suggested herein.

The *Guidelines* conclude with a brief section on *Additional Considerations* that cover the following topics:

- *Creation & Acquisition*
- *Preservation Partners & Permissions*
- *Distribution vs. Backups*
- *Change Management*
- *Preservation Monitoring*
- *Recovering Digital Newspapers from Preservation*

We have listed all tools and resources recommended in the *Guidelines* in the *Reference* section at the end. All of the recommended tools, with only one or two exceptions, are free and open source.

Roadmap Checklist

Essential Readiness		*Optimal Readiness*	
Inventorying			
	Hold Inventory Planning Meeting		Hold Inventory Planning Meeting
	Identify Collection File Locations		Establish an Inventory Workstation
	Create Simple Inventory Document		Identify Collection File Locations
	Record Basic Collection Information		Create or Use Existing Inventory Instrument
	Date Stamp & Version Inventory		Record Extensive Collection Information
			Create Checksums and UIDs(see below)
			Date Stamp & Version Inventory
Organizing			
	Analyze Current File/Folder Conventions		Analyze Current File/Folder Conventions
	Test GUI Tools for Changing File/Folder Names		Test CLI Tools for Changing File/Folder Names
	Apply New Conventions		Apply New Conventions
	Quality Check Modifications		Quality Check Modifications
Format Management			
	Identify File Format		Identify File Formats
	Record File Formats in Inventory (see above)		Store Batch Outputs and Update Inventory
	Document Acceptable Format Policies		Consult Format Registries
	Test Normalization/Migration		Document Acceptable Format Policies
	Normalize/Migrate Formats per Policy		Test Normalization/Migration
			Normalize/Migrate Formats per Policy
Metadata Packaging			
	Identify Metadata File Locations		Identify Metadata File Locations
	Record Metadata-to-Object Linkages		Record Metadata-to-Object Linkages
	Export/Package Metadata Records		Export/Package Metadata Records
	Store Metadata with Collection Files		Normalize and Consolidate Metadata to XML
			Create Preservation Metadata
			Store Metadata with Collection Files
Checksum Management			
	Create Per-File Checksums with BagIt		Create Per-File Checksums
	Routinely Audit w/ Stored Checksums with BagIt		Backup Checksums and Update Inventory
			Routinely Audit w/ Stored Checksums
Packaging			
	Document Location of Preservation Copies		Implement METS and/or PREMIS
	Document Preferred Recovery Media/Methods		Assign Unique IDs (UIDs) and Updating Inventory
	Document Recovery Process		Use UIDs with METS and/or PREMIS
	Document Access Copy Reproductions		Package Collection
	Store Documents with Collection Files		Package Collections together
	Package Collection		
	Package Collections together		

Section 1. Inventorying Digital Newspapers for Preservation

Rationale

Identifying what digital newspaper content an institution possesses is the first step in understanding its current and future preservation needs. Digital newspapers are created and acquired by institutions under a diverse array of circumstances, over wide spans of time, and often under the care of multiple managers. For that reason, an institution's digital newspaper collections and corresponding content files may reside on a range of different storage media (external hard drives, CDs, disk arrays, tape systems, etc.) and in multiple locations.

Identifying the amount (number of files and sizes), type(s), and location(s) of an institution's digital newspaper collections is a critical first step in preparing those collections for long-term preservation and archival management. An inventory can help curators not only determine where content resides, but what sorts of information may be needed to ensure its sustainability over time (e.g., format identification, checksums, digital object identifiers, etc).

Sound Practices

Producing a digital newspaper inventory is a multi-stage process that helps an institution establish what newspaper content it holds and where that content resides. Good inventories range from general to detailed, depending on the needs, goals, and resources of an institution.

An institution's digital newspaper inventory may be created and maintained as a text document, spreadsheet, or database. It should be easy to use, available to curators and technologists, scalable for future growth, and updated regularly. It should also explicitly record a "last updated" date.

A basic inventory describes characteristics of an institution's full range of newspaper files, including title, content type, format, size, associations (title/issue and page/article information), and file location. A more detailed inventory might also include information about creation date, copyright/permissions, software needed to render objects, checksums, and digital object identifiers.

Tools

The inventory may be maintained as a simple text document, spreadsheet, or database depending on the institution's needs and abilities. There are various platform-specific file managers such as *Nautilus*[24] for Linux, *Finder*[25] for Mac, or *File Explorer*[26] for Windows, to name just a few, which can be helpful for those less familiar with command-line approaches.

[24] The GNOME Project, "Nautilus," available at: https://wiki.gnome.org/action/show/Apps/Nautilus.

[25] Apple, "Mac OS X: Finder Basics," available at: http://support.apple.com/kb/VI209.

[26] Microsoft, "How to work with files and folders," available at: http://windows.microsoft.com/en-us/windows-8/files-folders-windows-explorer.

For those with more technical experience, command-line programs can provide a powerful pathway to begin inventorying digital newspaper collections. For example, a local technical staff member or system administrator may use basic Unix file commands (e.g., *ls*[27], *find*[28], *locate*[29], and *file*[30]) combined with shell scripts[31] to analyze a directory listing of files and output the information in a tabular format that can be modified in a spreadsheet (e.g., *TXT*, *TSV*, or *CSV*). The various output files can then be combined and organized in report documents as needed by a curator.

Readiness Spectrum

A digital newspaper inventory may be basic or detailed, and typically, the level of detail recorded in an inventory reflects the level of preservation an institution supports. As an institution matures in its lifecycle management processes, its inventory may likewise mature to include additional information.

No particular preservation readiness step - including the inventory creation - happens in isolation. Instead, institutions will circle back around to update their inventories as other preservation readiness activities occur. To this end, institutions at early stages of preservation planning should produce inventory instruments that will scale up over time to include additional categories of information. Institutions should also think carefully about categories of information that are likely to change, including file location (e.g., as an institution upgrades its storage media) and file type (e.g., if migrations or normalization occur).

At the lower end of the preservation readiness spectrum, an institution might start small, but build a strong foundation for future work. Simply establishing a flexible inventory document and identifying and recording information about all of an institution’s digital newspaper collections at a basic level (e.g., collection file names, file sizes, file locations, and file types) is an important beginning. Some institutions will accomplish this work without systems experience, either through manual entry or very basic systems report generation. At the higher end of the preservation readiness spectrum, an institution likely will use automated mechanisms to gather this information, plan for the inclusion of future collections, and record additional information (e.g., checksums and digital object identifiers).

Essential Readiness

An institution should consider the inventorying process as an ongoing activity that will require institutional planning, clearly assigned roles, coordination between curators and technical staff (where applicable), and on-going analysis and quality control.

A key first step for institutions of all types and sizes is to begin a planning phase. For very small institutions and/or collections, one person may undertake this planning step alone. Larger

[27] LinuxQuestions.org, “ls,” available at: http://wiki.linuxquestions.org/wiki/Ls.

[28] LinuxQuestions.org, “find,” available at: http://wiki.linuxquestions.org/wiki/Find.

[29] LinuxQuestions.org, “locate,” available at: http://wiki.linuxquestions.org/wiki/Locate.

[30] LinuxQuestions.org, “file,” available at: http://wiki.linuxquestions.org/wiki/File_%28command%29.

[31] LinuxCommand.org, “Writing Shell Scripts,” available at: http://linuxcommand.org/lc3_writing_shell_scripts.php.

Reference: Inventory Readiness Checklist

1. Complete file count and sizes (MB/GB/TB/PB)
2. Complete list of associated title/issue folders and page/article files (with metadata and OCR)
3. File location paths (if not included in the above)
4. File format extensions (MIME types)
5. Checksums and hash-functions (if available)
6. Digital object identifiers and identifier schemes used (if available and/or deemed useful)

institutions likely will need to arrange a planning meeting that assembles the range of digital newspaper curators and stakeholders across the institution. The goal of this planning is to gather and share information about the range of current and legacy content and the processes that impact the acquisition and storage of digital newspaper content over time.

Even though an institution may face short-term barriers to inventory creation (e.g., a shortage of in-house technical expertise), a collection curator at almost any institution should be able to complete the following essential tasks in preservation readiness.

1. Identify the institution's digital newspaper collections
2. Produce a text document, spreadsheet, or database inventory document within which information about the collections may be recorded
3. Document basic information including newspaper titles, file counts, file locations (i.e., storage media), and file names. This may be accomplished manually if necessary with the help of various file manager programs like those mentioned previously in the *Tools* (e.g., *Nautilus* for Linux, *Finder* for Mac, or *File Explorer* for Windows, etc.)
4. Date stamp and record the version of the inventory document

This *essential* stage of the inventory process gives a curator the opportunity to experiment with inventory applications and formats that can later extend to accommodate the fuller range of information elements desired and enable the necessary access to and update of data over time.

Optimal Readiness

Institutions with more time, technical staffing, and support resources can invest more deeply in this crucial first stage of preservation readiness. Compiling a thorough and well-organized inventory of the institution's digital newspaper assets will serve the institution well throughout the subsequent/overlapping stages of preservation readiness.

Institutions with technical resources and distributed collections would be well advised to establish a reliable workstation that can provide a centralized channel to all existing digital newspaper data within the institution. For collections not currently maintained on spinning disc, this workstation can serve as a staging apparatus and a curation workbench (e.g., collections maintained on CD or external hard drive might be temporarily staged on the workstation to gather information to populate the inventory and perform other preservation readiness activities). This workstation can serve as an ongoing conduit for gathering information and producing systems-based reports of the newspaper collections and files.

If designated staff have the proper permissions and authority to perform checksum creation and to mint object identifiers, these tasks may be completed during the inventorying process. Please see *Section 5. Checksum Management for Digital Newspapers* for more details regarding file fixity checksum creation and management. *Section 6. Packaging Digital Newspapers for Preservation* discusses the importance and use of object identifiers.

An *optimal* level inventory would include all those elements mentioned above under *Essential Readiness* but would also provide a more exhaustive overview of the collections, including:

1. Changes that have taken place over time to collection files (i.e., normalization, migration, etc)
2. File format types and the applications and software platforms that created them
3. File fixity information (if created) as well as the hash-function algorithms used to create them (e.g., *md5*, *sha-1*, *sha-256*, etc.)
4. Digital object identifiers that have been assigned (e.g., *ARKs*, *Handles*, etc.)

For institutions at mature stages of preservation readiness, much of this information may already be available in various locations and could be retrieved and incorporated into a consolidated inventory. For institutions that have not yet engaged with deeper preservation readiness activities, this information will flow out of some of the additional activities mentioned in later sections - namely *Section 3. Format Management for Digital Newspapers*, *Section 4. Metadata Packaging for Digital Newspapers*, and *Section 5. Checksum Management for Digital Newspapers* - and can be consolidated at that time.

Getting Started

The California Digital Library (CDL) has a well-cited example of what is called a Digital Assets Submission Inventory. This sample inventory document is a resource that can serve as a template for institutions seeking a simple means to build out an inventory. Among other things it seeks to guide the institution toward identifying file types, file sizes, number of objects, types of metadata, and types of storage media related to a collection or set of collections.

The document is available at: www.cdlib.org/services/dsc/contribute/docs/submission.inventory.rtf.

Institutions seeking to inventory their digital newspapers can adapt these simple tables to more accurately record information in-line with the suggestions recommended above.

Section 2. Organizing Digital Newspapers for Preservation

Rationale

Organizing digital newspaper content is a process through which an institution assesses, documents, and sometimes refines its file naming and folder usage conventions.

As mentioned in the previous section, an institution's newspaper content is often created and/or acquired by a range of players and over a long span of time. Different collections within an institution's holdings may conform to different file-naming conventions and folder conventions. Documenting these conventions clearly and/or normalizing these disparate collections by applying a unified schema enables future curators (and users) to retrieve, validate, and if necessary, reconstitute these collections in the future. As such, organizing digital newspapers is an important step in the preservation readiness process.

This process of organizing digital newspaper content builds upon and may go hand-in-hand concurrently with the Inventorying work described in the previous section as well as some of the additional preservation readiness activities covered later in the *Guidelines*, specifically:

- Inventorying the amount and location of digital newspaper content an institution is managing;
- Identifying the range of file formats and performing any necessary normalizations or migrations;
- Exporting and consolidating metadata for all collection(s); and
- Producing checksum manifests for this content.

Sound Practices

Sound practices for organizing digital news content primarily include the following:

- Rectifying any file-naming conventions that put content at risk of non-renderability;
- Documenting effectively the range of file and folder naming practices and conventions represented in an institution's collections; and
- Storing this documentation with the content it describes.

Even institutions with low resources and disparate practices will be able to provide a brief summary of each digital news collection's internal conventions that can help future curators and users understand each collection's structure for future use and renderability. Institutions with higher resource levels may also analyze and streamline their conventions and practices across digital newspaper collections.

The goal is to arrive at a documented and uniform approach (or small set of approaches) with clearly designated use-cases (e.g. one approach for legacy digitized content, another for recent digitization efforts, and a third for born-digital content) that contains clear guidelines for file naming and folder/sub-folder usage. Collection managers should coordinate with their technical staff members (or those of any external repository service provider) throughout any remediation and convention-setting process so that any change to existing conventions is understood and accounted for in the repository software environments used for access and/or preservation purposes.

Tools

Applying and enforcing a set of uniform folder and file-naming conventions can be a time-consuming endeavor if approached on a manual, per-file basis. There are some tools that can be used to batch rename and even relocate digital files. The best approach to take with these tools is to start with a cleanly copied representative subset of the overall data to which you would like to apply such tools. This will require setting aside a workspace where the tools can be installed and where the data can be copied to for testing purposes.

Examples of file-naming and re-locating tools include:

- Unix commands (e.g., *mv*[32] and *sed*[33])
- *GPRename*[34]
- *Bulk Rename Utility*[35]
- *Automator*[36]

Unix-based systems (including Mac OS X) come equipped with simple programs such as *mv* and *sed* that can be used to begin the organization process.[37] These programs can be used as standalones or in conjunction with various shell scripts to both batch rename and relocate files. A number of renaming tools with graphical user interfaces exist for most major operating systems. Among these, *GPRename* for Linux and *Bulk Rename Utility* for Windows offer a good range of features such as using regular expressions to find portions of a file name to replace, appending date information, and adding sequential numbers to files. *Automator* for Mac OS X has a graphical user interface, and despite some workflow limitations, can also assist with batch file renaming.

Readiness Spectrum

Organizing digital newspaper content can vary across a wide spectrum of practice while still fulfilling the basic goal of providing future curators with the information they need to understand the structure of each digital newspaper collection. This facilitates the curator's ability to preserve and render content reliably over time.

At the lower end of the preservation readiness spectrum, an institution may focus upon four core tasks:

- Identifying problems in file names that could compromise those files in the future;
- Using basic systems tools to perform batch renaming of these files (starting in a test-bed environment!);

[32] LinuxQuestions.org, "Mv," available at: http://wiki.linuxquestions.org/wiki/Mv.

[33] LinuxQuestions.org, "Sed," available at: http://wiki.linuxquestions.org/wiki/Sed.

[34] GPrename, "GPRename," available at: http://gprename.sourceforge.net/.

[35] Bulk Rename Utility, "Bulk Rename Utility Homepage," available at: http://www.bulkrenameutility.co.uk/Main_Intro.php.

[36] Apple, "Mac Basics: Automator," available at: http://support.apple.com/kb/ht2488.

[37] Note that there can be differences in syntax between GNU and BSD distributions. Mac OS X ships with BSD.

Case Study: File Naming Conventions

Below are some real-world examples of file name conventions that do a good job of providing title, issue, date and other unique id encodings. They include examples from both digitized and born-digital newspaper collections. They are just examples not standards.

Digitized Newspaper Examples (*PDF*, *TIF*, and *JP2*)

- 051-AAR-1873-09-24-001-SINGLE.pdf (title code/date)
- DCC_19601125-19600101_DLH_217.tif (title code/dates)
- bcheights_20040406_0001.jp2 (title code/date)

Born-Digital Newspaper Examples (e-print and web)

- an970607.pdf (title code/date)
- morning_725_5977.html (Morning Ed/7:25am/May 9, 1977)

Born-digital e-prints and web files often use the same filenames and extensions for both preservation and access copies. Make sure changes to preservation copies adhere to their current access copy filename conventions.

- Documenting institutional conventions in a text document to help future curators understand the collection logic; and
- Updating the digital news inventory to reflect all changes.

At the higher end of the preservation readiness spectrum, institutions may also streamline file-naming and folder usage practices into one or more well-documented and unified convention(s). Institutions may then use batch tools to remediate content according to their chosen convention(s). After completing this remediation work, institutions should always update their inventories.

Organizing digital news content ultimately makes collections intelligible and recoverable in the short and near term. With that in mind, the goal of this activity is to refine and communicate collection structures, file identifications, and other relationships so that curators and preservation partners can care for these collections. There are both machine-based and human-based approaches that can be taken across the readiness spectrum to achieve these goals.

Essential Readiness

As described above, organizing collections for preservation should begin with an analysis of file-naming conventions.

File-naming conventions for digital newspapers should follow established good practices (documented below), including attending to those specific to digital news content. Examining and adjusting filenames prior to preservation action is imperative because many repository systems (both preservation-oriented and access-oriented) may refuse to handle content that does not conform to standard practices. At best, in these cases the files will not render properly. At worst, poorly named files will not be able to be ingested into and preserved in a repository at all.

General good naming practices include:[38]

- Avoiding the use of special characters in a file name. \ / : * ? " < > | [ ] & $,;
- Using underscores instead of periods or spaces;
- Avoiding lengthy names;
- Including all necessary descriptive information independent of where it is stored;
- Including dates and formatting them according to international standards (YYYY_MM_DD or YYYYMMDD);[39] and
- Including a version number on documents to more easily manage drafts and revisions.

Good practices for applying consistent folder and file naming conventions to digital newspaper content more specifically include:[40]

- Retaining any repository system-defined folder naming conventions if supplied - this can be helpful for restoring collections to those systems at a later date;
- Following a simple title, year, volume, issue, month, day schema for folder and sub-folder conventions;
- Identifying the title in the file name;
- Including the year, month, and date of the issue publication in the file name;
- Including the page or article sequence number in the file name when appropriate;
- Including the corresponding newspaper section name where helpful; and
- Including the correct file extension with each file (e.g., *TXT, PDF, TIF, JP2*, etc.).

Depending upon the number of digital news files an institution is managing and how many of these are problematic, rectifying filename problems may be done "by hand" (on a small scale) or through the use of software tools. Such tools as those mentioned above allow for batch renaming of files, so that if there is a regular problem (e.g., a space or special character that needs to be replaced collection-wide), this can be dealt with simultaneously across a large number of files. Be sure to thoroughly test tools and batch processing prior to implementation. Wherever possible, create a copy of each collection that needs attention and work with those copies to ensure that accidental damage is not done to the originals as these file-name problems are corrected.

This renaming process, including the tools used, should be documented and this documentation should be included with the collection upon packaging (see *Section 6. Packaging Digital Newspapers for Preservation*).

After an institution addresses potential gaps in its file-naming conventions, it can begin analyzing and documenting its overall collection structures, including folder and sub-folder usage. For institutions with limited resources, this may simply mean creating a text-based document that explains the collection structures as they currently exist and what data

[38] State Archives of North Carolina, "Digital Records Policies & Guidelines: Filenaming," available at: http://www.ncdcr.gov/archives/ForGovernment/DigitalRecords/DigitalRecordsPoliciesandGuidelines.aspx#filenaming.

[39] A description of the rationale and usage of the international date standard is available at: http://en.wikipedia.org/wiki/ISO_8601.

[40] Examples of file and folder naming conventions for digital newspapers being provided are derived from analyses of digital newspaper collections as provided by the Chronicles in Preservation (2011-2014) project partners (http://metaarchive.org/neh).

Case Study: Boston College

Below is an example (not a standard) of a digital newspaper collection organization scheme as deployed by Boston College.

```
bcheights.........................................................................(collection title folder)
 |---2004.........................................................................(annual volume folder)
        |---04....................................................................(monthly volume folder)
              |---06..............................................................(daily issue folder)
                    |---bcheights_20040406.pdf
                    |---bcheights_20040406_0001.jp2
                    |---bcheights_20040406_0001.xml
```

elements, such as unique identifiers, are vital to preserving the file relationships.

Optimal Readiness

For institutions with more resources, devising and implementing a meaningful, consistent set of folder/sub-folder relationships and schemas across all digital news collections will improve the preservation outlook of these collections.

The degree of work involved depends upon an institution's practices-to-date. In some cases an institution may only need to review and refine its existing collection structures. In other cases, an institution may need to reorganize its digital news content entirely according to a newly designed set of consistent folder and file-naming conventions. The above Case Studies are some examples of how institutions with extensive experience in managing digital newspaper collections have organized their collections. Institutions are also encouraged to reference the *NDNP Technical Guidelines* as mentioned in the *Introduction*.[41]

Once a set of uniform collection structures and file-naming conventions have been established, curators and technical staff can work together toward implementation. This process should always begin by experimenting with copies of a sub-set of the collection and the batch renaming and/or relocating tools that are most appropriate for the institution's needs. Once the remediation process is tested thoroughly, implementation can begin, again ideally using a copy of the content rather than the originals. Remediation work, including the tools used, should be documented and this documentation should be packaged with the collection (see *Section 6. Packaging Digital Newspapers for Preservation*).

Caution: BagIt is discussed later in *Section 5. Checksum Management for Digital Newspapers*. It should be noted that if you have previously placed your data into bags, moving or renaming the files will invalidate the bag manifest. If this is the case it is advisable either to rebag the data or to make any organization changes before placing your data into bags.

[41] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants," August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

Section 3. Format Management for Digital Newspapers

Rationale

For more than a decade, newspapers have been digitized by an array of different institutions (libraries, commercial vendors, etc.) according to a variety of image, document and text output specifications. During this same timeframe, institutions have been acquiring "born-digital" newspaper content, including both e-prints (often through the *File Transfer Protocol (FTP)* or hard drive exchanges from a publisher to a library) and web-based files (often "harvested" using web-capture tools like *Heritrix*[42] or obtained via FTP exchanges). The resulting digital newspaper files come in a variety of flavors, including those typical for digitized newspapers (e.g., *TIFF*,[43] *PDF/A*,[44] *JPEG2000*,[45] *XML*,[46] etc.) and for "born-digital" newspaper contents (e.g., *PDF*,[47] various image, audio and multimedia formats *HTML*,[48] *XHTML*,[49] *CSS*,[50] *JavaScript*,[51] etc).

Sound Practices

In order to ensure the longevity of digital newspaper content, an institution must *identify* what file formats it manages, , *validate* these files according to their specifications, and normalize and/or migrate these files according to the institution's policy decisions. The processes of validation, normalization, and migration of newspaper files are used by institutions to ensure that newspaper content can be effectively rendered over time.

Identifying file formats is a first step in understanding what file types an institution is managing. As described previously, recording this information in an inventory (see *Section 1. Inventorying Digital Newspapers for Preservation*) will help an institution track a range of file format complexities it will need to

[42] Internet Archive, "Heretrix," available at: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[43] Adobe Systems Incorporated, "TIFF," available at: http://partners.adobe.com/public/developer/tiff/index.html.

[44] AIIM Standards Wiki, "PDF/A," available at: http://pdf.editme.com/PDFA.

[45] Joint Photographic Experts Group, "JPEG2000," available at: http://www.jpeg.org/jpeg2000/index.html.

[46] W3C, "XML 1.0," available at: http://www.w3.org/TR/REC-xml/.

[47] AIIM Standards Wiki, "PDF," available at: http://pdf.editme.com/PDFREF.

[48] W3C, "HTML," available at: http://www.w3.org/html/.

[49] W3C, "W3C XHTML2 Working Group Home Page," available at: http://www.w3.org/MarkUp/.

[50] W3C, "Cascading Style Sheets," available at: http://www.w3.org/Style/CSS/.

[51] Mozilla Developer Network, "JavaScript," available at: https://developer.mozilla.org/en-US/docs/Web/JavaScript.

address over time (including format obsolescence and format changes).

Format validation, briefly defined, is a process by which an institution assesses the conformance of a file to its format specification (e.g., that a *PDF* follows the internal content, layout, and structure rules of the *PDF* specification) and checks that a file will be rendered dependably by the programs designed for that format. Validating files allows an institution to catch and address errors (files that do not behave as they should).

Normalization is the process of converting numerous, diverse files from their native formats into a smaller number of more open, preservation-oriented formats, typically upon deposit or ingest (e.g., migrating articles transcribed through *OCR*[52] from Olive's *PrXML* to *METS-ALTO*). Migration more generally may be employed to ensure that the content of a file type that is facing obsolescence can be rendered into a new format (proprietary or open).

The library/archive communities have reached consensus regarding well-understood high-quality open archival formats for image-related collections like digital newspapers, namely *TIFF*, *PDF/A*, and to some degree even *JPEG2000* (lossless or lossy compression image format). The same cannot always be said for OCR and other article-level transcriptions, but curators and vendors typically aim to produce XML-based formats that have forward-migration pathways. Born-digital news (e-prints) may be contained in various legacy *PDF* and *HTML* versions, and web-based content (including social media content) may include a wide range of file formats, depending upon the particular born-digital newspaper.

Once an institution possesses a clear understanding of the range of different formats it hosts, it may determine that files need normalization or migration attention. The decision to normalize or migrate formats should be thoroughly evaluated. Consultation with one or more format registries is a first step - this can help an institution to identify potential migration pathways to more suitable formats. The institution can then familiarize itself with various tools that can perform necessary transformations.

We note that normalization and migration are still fraught topics in the library/archive realm, with passionate advocates both for and against employing these practices. We also note that format registries and migration tools are still in early stages of development, and should be used only after thorough consideration; it does not hurt to cross-reference these registries. In addition, format migration does not require, nor should it imply, that a content curator should dispose of the original or current format. It is advisable to continue preserving both the original and successor formats for as long as resources permit.

Tools

Content curators need lightweight tools to help them determine the full range of different file formats that comprise their digital newspaper collections and assess whether these file formats are valid according to their specifications. It should be noted that format identification and validation tools are limited in

[52] Library of Congress, "National Digital Newspaper Program: Digitizing Microfilm and Optical Character Recognition," December 2012, available at: http://www.loc.gov/ndnp/guidelines/digitizing.html.

the types of formats that they can reasonably identify and validate - in some cases multiple tools may be needed to validate outputs for a single collection. Some format identification and validation tools can also produce technical metadata (more on this in *Section 3. Metadata Packaging for Digital Newspapers*).

Helpful format identification and validation tools include:

- Apache *Tika*[53]
- *Digital Record and Object identification (DROID)*[54]
- *JHOVE2*[55]
- Unix *find*[56] and *file*[57] commands
- *File Information Tool Set (FITS)*[58]

Normalization and migration decisions are ultimately policy decisions. There is no "right" answer regarding whether or not these activities are necessary or advisable for a particular institution. In order to establish local policy, an institution should consider the following:

- Level of need: Does the institution have obsolete digital newspaper formats?
- The viability of the institution's current digital newspaper formats.
- The range of the institution's current formats: Is it so broad that the institution's ability to keep track of viability is compromised?
- Resource levels: Is it feasible for the institution to test and run any format management tool?

If normalization and/or migration are undertaken, the tools an institution uses should be thoroughly tested prior to implementation.

Helpful format registries include:

- Archiveteam *File Format Wiki*[59]
- Library of Congress *Sustainability of Formats*[60]
- *PRONOM*[61]
- *Unified Digital Formats Registry (UDFR)*[62]

[53] Apache Software Foundation, "Apache Tika," available at: https://tika.apache.org/. Apache Tika has a number of use cases. Libraries are just recently starting to benchmark its usage for performing file format identification – see here: http://www.openplanetsfoundation.org/blogs/2013-05-20-apache-tika-file-mime-type-identification-and-importance-metadata.

[54] UK National Archives, "Droid," available at: http://digital-preservation.github.io/droid/.

[55] "JHOVE2," available at: https://bitbucket.org/jhove2/main/wiki/Home.

[56] LinuxQuestions.org, "Find," available at: http://wiki.linuxquestions.org/wiki/Find.

[57] LinuxQuestions.org, "File," available at: http://wiki.linuxquestions.org/wiki/File_%28command%29.

[58] Harvard University Library Office for Information Systems, "File Information Tool Set," available at: http://project.iq.harvard.edu/fits.

[59] Archiveteam, "Archiveteam File Format Wiki," available at: http://fileformats.archiveteam.org/.

[60] Library of Congress, "Library of Congress Sustainability of Formats," available at: http://www.digitalpreservation.gov/formats/.

[61] UK National Archives, "PRONOM Technical Registry," available at: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

[62] California Digital Library, "Unified Digital Format Registry (UDFR)," available at: http://udfr.cdlib.org/.

Helpful format normalization and migration tools include:

- *Adobe Image Processor*[63]
- *ImageMagick*[64]
- *Xena*[65]

Readiness Spectrum

The complexities involved in managing formats for digital newspapers will vary depending upon two main factors:

- The range of formats an institution holds
- The institution's decisions regarding normalization and migration activities

Institutions with consistent digital newspaper collections that include a small number of formats will find preservation readiness work easier than institutions with inconsistent collections that cover a wider spectrum of formats. And institutions that do not engage in format migration/normalization activities (whether due to resource or policy-based decisions) will find format management less time/resource intensive, at least in the short term.

Initial management steps vary little across the broad readiness spectrum. Almost any institution will be able to complete the following tasks:

- Identify and document its digital newspaper file formats using tools like *DROID*, which has a graphical user interface (GUI) and links to *PRONOM*
- Evaluate and make determinations about sustainability issues presented by the various identified formats (using the Archiveteam *File Format Wiki*, *UDFR*, *PRONOM*, and/or the Library of Congress *Sustainability of Formats* website)
- Establish policies regarding normalization and migration
- Normalize or migrate files deemed "at risk" (e.g., using tools like *Adobe Image Processor*, *ImageMagick*, or *Xena*)

More advanced institutions may:

- Identify, validate and normalize/ migrate formats by using command-line oriented tools such as Apache *Tika*, *JHOVE2*, *FITS*, Unix programs like the *find* and *file* commands (or their corollaries in other OS environments) combined with shell scripts and *ImageMagick*.

The essential and optimal steps are described in more detail below.

Essential Readiness

Institutions with limited resources can fulfill the *essential* steps in preservation readiness by documenting the formats they are managing and the potential sustainability issues associated with these formats. If the institution is managing reliable formats (i.e., formats that are not obsolete or in danger of obsolescence in the near-term), a regular process of re-checking and carefully documenting the file formats in its collections can suffice for a basic preservation

[63] Adobe Systems Incorporated, "Image Processor Script," available at: http://tv.adobe.com/watch/understanding-adobe-photoshop-cs6/image-processor-script/. Adobe's software requires a paid license, but it is widely used for bulk image format migrations.

[64] ImageMagick Studio LLC, "ImageMagick," available at: http://www.imagemagick.org/.

[65] National Archives of Australia, "Xena – Digital Preservation Software," available at: http://xena.sourceforge.net/.

Case Study: Boston College

Boston College has digitized several of its campus newspapers in accordance with the *National Digital Newspaper Technical Guidelines*. This has provided Boston College with several high-quality archival page scans in both *TIFF* and *JPEG2000* formats.

To conserve storage space Boston College has opted to prioritize its *JPEG2000* images as preservation masters (*TIFF*s can be quite large). This retains the legibility of text and graphics. Due to the amount of white space they included, the images were eligible for some small amount of compression. While *JPEG2000* is not as widely adopted as *TIFF*, Boston College believes this will change and the format still satisfies the criteria for being non-proprietary and open source.

Boston College has also tested the conversion from *JPEG2000* back to *TIFF* with satisfying results.

readiness step in the short term. If the institution identifies obsolescence issues for any of the formats it manages, however, it should strongly consider migrating the at-risk files to a stable file format (remember that this does not need to lead to removing support for the original format).

File identification is the first step in format management, and there are a number of ways to fulfill this step. One lightweight way an institution with limited time or modest technical skills may accumulate basic knowledge about its file formats is to work with a technical staff person or system administrator to install a tool like *DROID* that has a graphical user interface (GUI) and has direct links to *PRONOM*, one of the longer-standing format registries. Once the institution understands the format types it holds and their associated risk factors, the institution may make policy-based decisions regarding what normalizing and migration activities it must take and what staffing/resources will be required or partnerships it will need to form in order to accomplish this further work. *Xena* (see above) is a well-documented format normalization tool that also has a GUI that should be relatively easy for an institution to begin working with.

Optimal Readiness

An institution with more time, expertise, and resources to expend should pursue a multi-step workflow to identify and address problematic formats in its collections.

Institutions with more technical staffing might prefer to use more advanced command-line approaches. Unix programs such as the *find* and *file* commands (or similar tools in other OS environments) can be used in concert with a shell script to create a per-file list of MIME type values at a top-level or sub-directory level. This list can then be exported to a tabular format (e.g., *TXT*, *TSV*, or *CSV*) for further analysis and format tracking. The institution can store this output file and/or any derivations (e.g., *XLS*, *TXT*, *DOCX*, *PDF*, etc.) in a sub-folder(s) along with the corresponding directory of analyzed files. Ideally, the directory name and date should be included in the filename(s) of this file(s). If files are added to the collection over time, the commands can be re-run, and a new set of outputs stored. Tools, such as *FITS*, go a step further to not only identify file formats but validate their conformance to the format. They can also provide report outputs in several tabular formats such as those mentioned above, as well as in *XML*.

Case Study: Virginia Tech

Virginia Tech has been a leader in working with publishers of born-digital newspapers to archive those publishers' *PDF*s and web files.

To better manage and preserve the born-digital web files under its care, Virginia Tech migrated early versions of this *HTML* content to the more recent *HTML 4.0*. Though this was a significant undertaking it enabled Virginia Tech to apply better consistency and reliability for the rendering of this unique content.

Through its participation in the NEH-funded Chronicles in Preservation project, Virginia Tech was also given the opportunity to apply leading format identification and validation tools such as *DROID*, *JHOVE2*, and *FITS*. These tools were especially helpful for characterizing its early versions of *PDF* content.

In addition, the Chronicles in Preservation project enabled Virginia Tech to make use of the *FCLA Description Service*[66] to generate technical metadata for its born-digital file formats.

Once the institution has this basic knowledge about the file formats it manages, it can explore and experiment with some of the nascent format registries to determine any sustainability issues that these formats may present (e.g., obsolescence, lack of open standards, backwards compatibility issues, etc.). An institution can conduct this research using the Archiveteam *File Format Wiki*, *UDFR*, *PRONOM*, and/or the Library of Congress *Sustainability of Formats* website. For example, an institution's analysis of its file formats may reveal several born-digital newspaper files in the *HTML 2.0* format. With this information, the institution could then turn to the *UDFR* and perform a search on *HTML 2.0* and return a full format profile, identify its successor format versions (in this case *HTML 3.2*, *4.0*, and *XHTML*), the applications that were able to output files into this format, and the applications that can successfully render *HTML 2.0* documents.

This institution can then set up a test-bed environment for experimenting with migration and normalization tools like *ImageMagick* or *Xena*. Using subset copies of its born-digital newspaper content the institution could experiment with *Xena*'s in-built features for converting this legacy *HTML* content into valid *XHTML*.

Once the format risk factors and the migration pathways are both known and thoroughly tested, the institution can make a policy decision regarding normalization and/or migration for files stored in this format. Depending upon the policy, the institution may choose to normalize, migrate, and/or continue to store its current format types.

Downloading, installing, and testing the various utilities and tools mentioned above will require work by technical staff, curators, or consultants with command-line experience. Structuring and

[66] This was accomplished using a script program known as bag-describe developed in the Chronicles in Preservation program. It invokes the FCLA Format Description Service to generate PREMIS technical metadata for all objects contained within a Bag. That script is documented and available at: https://github.com/MetaArchive/bag-describe. The hosted service is available at: http://description.fcla.edu/.

making sense of the outputs from such tools will also require some investment of time (and patience).

Finally, performing format migrations requires a larger resource investment than other format management steps. To perform migrations, an institution should ideally set aside a workstation with adequate space, processing capability, and configurations for a proper test-bed environment. The institution will need to determine and test the proper migration tools, a task that will necessarily involve both curators and staff with experience in installing and configuring open-source software. The institution should run sample conversions and perform manual quality checks prior to any batch migrations, and all migrations should be deployed in accordance with institutional policy documentation. Coordination between technicians and curators will be needed throughout the migration process.

Section 4. Metadata Packaging for Digital Newspapers

Rationale

Preservation readiness for metadata refers to the process of ensuring that metadata is properly preserved along with the item/collection it describes. Particularly in the case of digital newspapers - which are often compound objects with complex relationships - maintaining robust connections between the metadata and the content is essential. If this information about the objects is lost, they may no longer be able to be reassembled and used in meaningful ways.

In this guidance document, content curators will find practical advice regarding:

- General strategies for packaging metadata for digital newspapers
- Responsible approaches for exporting metadata from any existing repository systems where it may be held
- Strategies for navigating the features of such repository systems
- Tips for managing the outputs from all such activities

Please note: this section does not provide advice on collection-level descriptive metadata creation (e.g., *MARC*, *DublinCore*, etc.). Institutions needing advice on that topic should refer to the host organizations for the various standards and schemas that may apply. This section does address the importance of creating administrative, technical, and structural metadata for long-term preservation purposes.

Sound Practices

Packaging metadata essentially requires knowledge of three core factors: 1) the metadata an institution currently has for its newspaper content; 2) where this metadata is stored, and 3) how this metadata is related to the objects/collections it describes. Each of these factors will be considered below.

1. What metadata does an institution have?

Institutions produce metadata to aid with collection description, discovery, and archival management. Multiple schemas (e.g. *Dublin Core*,[67] *METS*,[68] *PREMIS*,[69] *MIX*,[70] and *MODS*[71] among others) are often used to record this descriptive, technical, administrative, and structural information for digital newspaper files and collections. Institutional practices vary

[67] Dublin Core Metadata Initiative, "DCMI Home," available at: http://dublincore.org/.

[68] Library of Congress, "Metadata Encoding Transmission Standard (METS)," available at: http://www.loc.gov/standards/mets/.

[69] Library of Congress, "Preservation Metadata: Implementation Strategies (PREMIS)," available at: http://www.loc.gov/standards/premis/.

[70] Library of Congress, "Metadata for Images in XML Standard," available at: http://www.loc.gov/standards/mix/.

[71] Library of Congress, "Metadata Object Description Schema," available at: http://www.loc.gov/standards/mods/.

widely in terms of schemas used over time, and also in the levels of completion or conformance to these schemas and their attendant profiles. Understanding and documenting what metadata schemas (and which versions) have been used for different collections and items over time is the first step in readying this metadata to be preserved with the content it describes.

2. Where is metadata stored?

The storage and maintenance of this metadata likewise varies widely. Most often, the metadata is stored either alongside the collections it describes (e.g., in a repository system as some type of associated file) or embedded with the objects/collections it describes (e.g., via *METS* or *METS-ALTO*[72] packaging or in file headers). Sometimes, metadata may also be held in a collection database that describes many digital collections held by an institution (e.g., one catalog that describes the entire digital holdings of an institution). An institution should document these locations within its digital newspaper inventory to ensure that curators know where the metadata resides (see *Section 1. Inventorying Digital Newspapers for Preservation*).

3. How is metadata related to the objects/collections?

There is a wide range of practices for metadata association and linkages. At the lower end of the spectrum, some institutions document these relationships through maintaining a metadata spreadsheet or database that includes keys or unique identifiers that correspond to each collection title or collection item. Some institutions fall back on their repository systems (e.g., *DSpace*,[73] *Fedora*,[74] *Olive ActivePaper*,[75] or *ArchivalWare*[76]) to structure both their collections and their metadata according to the repository software's default means for associating the records with the objects. Still others package the object with its metadata or refer to it externally using *METS* or another packaging standard.

Understanding the local range of metadata schemas, locations, and relationships is the first step in readying the metadata for preservation along with the objects/collections it describes.

Tools

There are a number of tools and approaches, ranging from simple to sophisticated, that can be of help in packaging metadata for preserving digital newspapers.

The first is rather straightforward and relates to the preservation readiness activity of inventorying covered earlier in the *Guidelines*. For institutions of all sizes, but particularly smaller or under-resourced institutions, digital newspaper content may have been acquired in

[72] Library of Congress, "ALTO Technical Metadata for Optical Character Recognition," available at: http://www.loc.gov/standards/alto/techcenter/use-with-mets.php.

[73] DSpace, "DSpace Homepage," available at: http://www.dspace.org/.

[74] Fedora, "Fedora Repository," available at: http://fedora-commons.org/.

[75] Olive Software, "Active Paper Archive," available at: http://www.olivesoftware.com/activepaper-archive.html.

[76] ArchivalWare, "Libraries," available at: http://www.archivalware.net/libraries.

ad-hoc ways and on a wide range of media, particularly born-digital newspaper content (e.g., pre-prints or web files from local presses and publishers). As such, metadata may reside in multiple locations and conform to a range of standards (or even no standards at all). The inventorying stage provides an opportunity to record where this metadata lives relative to the actual collection files. This can be done in an inventory instrument, or in a simple spreadsheet or a database used just for the purposes of metadata tabulation. The most important task is recording the associations between the metadata and the collection and/or items. More on what an institution should do with the outputs of such approaches in the next section on *Essential Readiness*.

For institutions that use repository software systems for their digital newspapers, as mentioned above, metadata often is stored within these systems in some relationship to the collection. This metadata may include various administrative, technical, and even structural elements. More often than not this metadata contains collection- or item-level descriptive information. The process of extracting and packaging any such metadata quite often falls to the native export features of the various repository systems being used to serve out the associated content. For example, if an institution stores its metadata in one of the popular repository software systems (e.g., those mentioned above: *DSpace*, *Fedora*, *Olive ActivePaper*, *ArchivalWare*, etc), the software may provide metadata export functions that can output the records as *XML*. Institutions should consult the system documentation and make use of developer support during this process to ensure consistent and thorough outputs that include the full range of metadata elements that should be derived from the system. Depending on the purpose of preservation it may be important to retain certain data elements (e.g., *Handles*[77] used in *DSpace*) for the sake of rebuilding the collections at a later date in the same repository environment. In other cases, this data may be extraneous to the long-term goal of the institution's preservation use case, and may be excluded accordingly.

Institutions producing digital newspaper collections according to well-formed digitization standards, such as the *NDNP Technical Guidelines*,[78] will already have adequate-to-excellent metadata records, including page-level *METS* or *METS-ALTO* records containing descriptive, technical, administrative, and structural metadata. In these cases metadata may have been produced by a vendor or digitization unit and packaged with the collection files according to those well-formed specifications. Such institutions may require very little additional work to consolidate their metadata for long-term preservation.

Institutions with adequate resources that have not yet moved beyond descriptive metadata for their digital newspaper collections have access to a range of tools that can assist with analyzing image, text, and other multimedia files, as well as extracting metadata from them for long-term archival management and metadata packaging. The *XML* outputs of these tools can then be

[77] Corporation for National Research Initiatives, "Handle System," available at: http://www.handle.net/.

[78] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants," August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

consolidated and conformed into schemas such as *METS* and/or *PREMIS*.

Below is a list of tools with varying sets of use cases and requiring different degrees of knowledge about file specifications:

- Unix *file*[79] command
- *New Zealand Metadata Extraction Tool*[80]
- *Exiftool*[81]
- *File Information Tool Set (FITS)*[82]

For example, the Unix *file* command can produce application and MIME type specific information on a per file basis and can be combined with shell scripts to recursively perform batch outputs, resulting in tabular formats that can be processed further for the sake of more long-term supported metadata schemas.

The *New Zealand Metadata Extraction Tool* is a Java-based graphical user interface (GUI) application that can run on Windows and Unix platforms to analyze files and output findings to XML.

Exiftool is a command-line utility that can read, edit, and create metadata for a wide variety of file formats.

FITS is an open source, command-line tool for Unix-based systems that combines the abilities of many different open-source file identification, validation, and metadata extraction tools and that outputs results to *XML*. The *XML* does not conform to any formalized metadata standard, but the output is well formatted and can be cross-walked to *METS* or *PREMIS* as necessary.

Readiness Spectrum

Under the best of circumstances metadata should be created according to specific collection needs but unified across collections wherever possible (e.g., to facilitate greater discovery and/or improve an institution's ability to manage the content efficiently and effectively across their repository environment). Institutions should choose metadata schemas and approaches that meet focused use cases, and create metadata that they can functionalize and sustain. The range of metadata managed by an institution can often be complicated by legacy approaches. Over time, different curators at different moments in time may have made different choices about metadata approaches, leading to institutional inconsistencies across content sets. As an institution begins to ready its collections for long-term preservation, it can impose consistency through mapping its metadata into a common format and enriching it for preservation using the tools described above. As the tools continue to improve, this remediation will come into reach for a broader range of institutions, including smaller and less resourced institutions.

[79] LinuxQuestions.org, "file," available at: http://wiki.linuxquestions.org/wiki/File_%28command%29.

[80] National Library of New Zealand, "Metadata Extraction Tool," available at: http://meta-extractor.sourceforge.net/.

[81] Phil Harvey, "Exiftool," available at: http://www.sno.phy.queensu.ca/~phil/exiftool/.

[82] Harvard University Library Office for Information Systems, "File Information Tool Set," available at: http://project.iq.harvard.edu/fits.

Case Study: Georgia Tech

Georgia Tech participated in the NEH-funded Chronicles in Preservation project. They exported three digital newspaper collections from their *DSpace* repository to package them with *BagIt* and send them to project preservation partners.

They ensured that the *DSpace* assigned *Handle* ID for each object was exported and stored. These *Handle* IDs are essential for future recovery.

Essential Readiness

For institutions with little time, resources, and available expertise to consolidate or enrich a set of diverse legacy schemas across their digital newspaper collections, the most crucial step is to document existing metadata practices:

- Identify all existing metadata (note the standard and its version)
- Establish the relationship between each metadata record and the digital newspaper content that it describes
- Document relationships clearly in an inventory document, spreadsheet, or collection database
- Package the inventory in a tabular format (so that it can be accessed by multiple programs)
- Store packaged inventory files with the content that it references, or in a readily identifiable folder

When metadata is more consistent and/or supported by software systems, institutions can improve their readiness through exporting all existing metadata and packaging it in a non-proprietary *XML* format (if not already done), and then storing such records either in a readily identifiable folder, or in a folder with the content it describes. Institutions should also make sure that each metadata record has and retains identifiable linkages to the collection content it describes through the use of unique identifiers or other regularized mechanisms (repository software systems often assign such information).

If an institution is assembling collections from multiple units or from multiple locations/ systems, it must ensure that unique identifiers will not conflict across collections. This can be accomplished by adding a unique collection or repository id as a prefix to every file id.

Optimal Readiness

Optimal readiness involves remediating and enhancing metadata prior to the preparation of a *Submission Information Package (SIP)* for preservation. It also may involve the use of *METS* or other schemas that embed the metadata with the content it describes.

For those institutions that have generated *METS* and/or *PREMIS* metadata associated with their digital newspapers at or after the time of creation, metadata may be either directly incorporated into those extensible schemas or should point to the locations where the various associated metadata resides. If the metadata is held within the *METS* and/or *PREMIS* records, these records should be retained in their proper locations in relationship to the collection files. Depending on the degree of connection between referenced metadata records and the collection data, efforts may be needed to retrieve the external records, store them logically alongside the collection data, and ensure that the identifier linkages to that data are accurate.

When dealing with HTML-encoded metadata for digital newspapers it may be worthwhile to pause before packaging and transferring web files to make sure that any <meta> tags are well-formed and that information will not be lost due to invalid *HTML*. Similarly, it can be helpful to make note of the use of any <pre> tags that may be used to display information - an element that is presumably not supported in upcoming versions of *HTML* (e.g., *HTML5*). Perhaps this metadata, because it is subject to a wide range of browser support dependencies, should be extracted and saved as separate records.

Institutions with available time and expertise to consolidate and enrich their metadata should make use of the most open, lightweight, and proven tools and make every effort to automate the process. The end goal is to create a series of digital newspaper objects with adequate technical, administrative, and - where necessary - structural metadata. Using some of the tools mentioned above, institutions can record information about file formats (technical) and the applications that created them (administrative). They can combine this information with the descriptive metadata in *METS* records - perhaps even leveraging the *METS* structural elements (and *METS-ALTO* where applicable) as well as incorporating checksums (see *Section 5. Checksum Management for Digital Newspapers*).

Section 5. Checksum Management for Digital Newspapers

Rationale

Stewards of digital newspapers need to be able to attest to the completeness, correctness, authenticity, and renderability of their collections over time. One way that institutions can do this is to require that checksums (digital signatures) be generated for their master digital files at the time of their creation and to store and compare these checksums over time. Recent digitization specifications and standards recommend that when institutions outsource digitization, they request listed records (or manifests) of files and their checksums from their vendors or digitization units and actively use these to verify their digitized collections upon receipt (i.e., to make sure that collections arrive intact). With this manifest of files and their checksums, stewards can also perform routine audits and implement repairs on corrupted objects from backups or preservation copies as needed over time.

Sound Practices

Checksums can be generated by several open source tools and utilities (more on this below), and can be stored in a simple txt file alongside the object, in an associated metadata object (e.g., *METS*), in a manifest with many other checksums, in a database, or any combination of the above. Once stored, these checksum records can be called upon by both content curators and preservation service providers to ensure that the objects have survived intact through both network based transfers and hardware/software processes.

When recording checksums for master digital newspaper files, a few important practices to follow include the following:

There are a several different kinds of checksum algorithms available for institutions to apply to their files (*md5*,[83] *sha-1*,[84] and *sha-256*[85] being the most prominent). Recording which algorithm was applied is imperative so that later verification processes can properly apply that same algorithm.

Checksums are data and subject to the same risks as any other data. To guard against potential misplacement, loss, or damage, it is good practice to keep backups of checksums.

Tools

There are many open source tools and utilities for creating and working with checksums for digital newspapers.

Examples of checksum creation tools include:

- *md5sum*[86]
- *sha1sum*[87]

[83] A description of the md5 algorithm is available at: http://en.wikipedia.org/wiki/md5.

[84] A description of the sha-1 algorithm is available at: http://en.wikipedia.org/wiki/sha1.

[85] A description of the sha-2 algorithm, of which sha-256 is a part, is available at: http://en.wikipedia.org/wiki/sha2.

[86] Linuxquestions.org, “Md5sum,” available at: http://wiki.linuxquestions.org/wiki/Md5sum.

- *hashdeep*[88]
- *Fixity*[89]
- *BagIt*[90]

Md5sum and *sha1sum* are standard Unix command-line programs and are usually invoked on a per-file or folder basis. These can be coupled with options and arguments for outputting results to various formats (*TXT*, *XML*, *CSV*, [91] etc.).

Hashdeep is a lightweight open source application that provides technicians with features and commands for creating and comparing checksums for digital objects at both file and batch levels. It includes a reporting function that explains the reason for a comparison test's failure.

Fixity is a lightweight program to automate checksum monitoring. Curators choose a regular interval with which the tool will generate checksums and compare values. Upon completion, reports are emailed to the curators.

BagIt is a packaging and transfer specification that can be applied to an existing set of organized collection content. As a specification, *BagIt* defines a data model called a bag that includes a folder with all the collection data and a manifest of per-file checksums and file pathnames (see the *Boston College Case Study* below for an example of using *BagIt* for checksum management). There are a number of existing tools that can create bags, such as *Bagger*[92] and *bagit-python*.[93]

Readiness Spectrum

Creating checksums for digital newspaper content is a relatively easy process. For small collections and non-technical environments (e.g., institutions without technical staff members), there are tools that can be used to calculate checksums. For example, graphical user interface (GUI) tools such as *Bagger* and *Fixity* can make batch checksum creation very easy and provide the institution with a ready-made manifest of files and checksums. The command-line programs mentioned above are relatively simple to invoke, and technical staff with even a moderate level of experience in Unix and Linux environments should have no problem coordinating the programs with scripts (or using tools like *hashdeep*) to automate the batch creation of checksums for multiple objects. Others that are more platform-specific can also be used.

Managing the checksums you have created requires additional effort. A checksum is only helpful when it is used consistently over time. Applications will need to create new checksums on demand and either automatically compare them back against previously recorded checksums or output them into a format that

[87] Linuxquestions.org, "Sha1sum," available at: http://wiki.linuxquestions.org/wiki/Sha1sum.

[88] Jesse Kornblum, "md5deep and hashdeep," available at: http://md5deep.sourceforge.net/.

[89] AVPreserve, "Fixity," available at: https://github.com/avpreserve/fixity.

[90] Descriptions of the BagIt specification are available at: http://en.wikipedia.org/wiki/BagIt and http://tools.ietf.org/html/draft-kunze-bagit-10.

[91] Wikipedia, "Comma-separate values," available at: http://en.wikipedia.org/wiki/Comma-separated_values.

[92] Library of Congress, "Bagger," available at: http://libraryofcongress.github.io/bagger/.

[93] Library of Congress, "bagit-python," available at: http://libraryofcongress.github.io/bagit-python/.

Case Study: Boston College

In the NEH-funded Chronicles in Preservation[94] project, several digital newspaper curators experimented with *BagIt* to inventory, checksum, and package their collections for preservation purposes.

Boston College packaged 183 GB of digital newspaper content using *BagIt*. This package was then split into smaller 30 GB archival units for preservation storage using the *BagIt Java Library*. These smaller archival units then had their checksum manifests validated against the original manifest using custom scripts built in the project.[95]

After a successful ingest into the MetaArchive preservation network, these smaller *BagIt* units were exported, rebuilt, and validated as the original 183 GB package using some additional custom scripts built in the project. The rebuilt *BagIt* package was then returned to Boston College who were able to validate checksums using the *BagIt* tools.

can be processed for comparison purposes such as a comma-separated values (*CSV*). Because different tools record checksums in different ways, it is important to be able to access and migrate the data in case the tool you are using is abandoned or better tools are developed. Additionally, the algorithm used to create the checksums must be recorded, and it must also be supported by the applications that will be used to create them and/or compare them in the future.

Depending on the scale of the content being managed, and the degree of sophistication in tools and approaches being used, an institution may want to document up-front its checksum management workflows in the context of its current or prospective data management environment (more on this next under *Optimal Readiness*).

Essential Readiness

Using *Bagger* is an excellent way to keep checksum information closely associated with the content for which it was created. The README instructions included with Bagger are easy to follow and allow staff to begin creating "bags" of collection content that will include a file called manifest-md5.txt or manifest-sha256.txt. The manifest file indicates the algorithm used to produce the checksums and lists the file-path and checksum of every file in the collection. Bagged content can be routinely validated with the *BagIt* tools. The command-line versions of the *BagIt* utilities have options and arguments that can be invoked to perform specific operations such as checking for missing files and validating checksums. *BagIt* is a specification widely adopted and used in the memory community. A bag created at one institution can be transferred and validated at another with ease.

[94] MetaArchive Cooperative, "Chronicles in Preservation," available at: http://metaarchive.org/neh/index.php/Main_Page.

[95] The scripts being referenced here were produced as a series of Interoperability Tools now available on GitHub at: https://github.com/MetaArchive.

AVPreserve's *Fixity* tool provides another easy way to create and monitor checksums over time. An institution with little to no access to technical staff or expertise can follow the User Guide and choose to validate checksums on a daily, weekly, or monthly schedule. A manifest listing the file-path and checksum of every file in the collection is stored as a *CSV* in the program's directory. When the tool runs, it generates new checksums and compares the values to those in the manifest. After the validation, an email is sent to up to seven recipients reporting whether files have been added, renamed, deleted, or corrupted.

Optimal Readiness

For institutions with more time, resources, and expertise, command-line programs like *md5sum* or *sha1sum* will provide more flexibility regarding the application of the task and control over the output format (*BagIt* gives you a quick solution but has some mild dependencies on the *BagIt* utilities). As mentioned previously, there are also versatile tools such as *hashdeep* that can facilitate batch creation of checksums, and that provide a suite of features for comparing checksum digests.

After checksums have been created, they must be properly managed over time. An institution should store its checksums in secure locations, developing logical schemas and approaches for associating checksums to the digital newspaper objects for which they were generated, and establishing reasonable schedules and workflows for creating checksums and comparing them back against their previously generated counterparts. Because checksum validation is a task that must be performed regularly, it should be automated as much as possible, whether with operating system features such as *cron*[96] in Unix environments and *scheduled tasks*[97] in Windows or as part of a larger repository environment. Establishing regular audit and reporting schedules and enforcing these within the institution's broader digital preservation policy helps to ensure that the practice is carried out routinely over time. Audit schedules should be logical and take into consideration the overall amount of data that needs to have checksums generated and compared at any given interval (checksum creation and comparison operations, particularly when involving large amounts of data, can be time and CPU intensive).

[96] Linuxquestions.org, "Scheduling tasks," available at: http://wiki.linuxquestions.org/wiki/Scheduling_tasks

[97] Microsoft, "Schedule a task," available at: http://windows.microsoft.com/en-US/windows/schedule-task.

Section 6. Packaging Digital Newspapers for Preservation

Rationale

The previous sections of these *Guidelines* have been geared towards preparing digital news collections and content for packaging. Whether an institution is seeking to store digital news content locally or to exchange its digital newspapers with an external preservation service provider, properly packaging this content provides curators with the necessary information, controls, and linkages to manage digital newspapers over time, including through changes in storage media, while still maintaining the integrity of both the objects and collections.

In formal *OAIS* terms, "packaging" is the process of establishing a *Submission Information Package (SIP)*[98] for deposit in a preservation repository.

Packaging can be accomplished in multiple ways, including:

- Using simple (text-based) documentation strategies;
- By deploying unique identifiers (*UID*s) coupled with Name Assigning Authorities and other local metadata and management tools; and/or
- Via lossless packaging formats like *TAR*,[99] *WARC*,[100] or self-describing preservation specifications such as *BagIt*[101] (among other approaches).

Using lossless packaging formats helps to hold content together as it is moved into archival storage, requested for routine audit purposes, and/or being sent to and from preservation partners over time. Each of these elements will be covered in the sub-sections ahead.

Sound Practices

The *Reference Model for an Open Archival Information System (OAIS)* defines the range of information elements that assist in long-term preservation and access as *Preservation Description Information (PDI)*. The *Reference Model* also describes three information package models:

- *Submission Information Packages (SIPs)*
- *Archival Information Packages (AIPs)*

[98] See *Reference: OAIS Packaging Concepts & Models* later in this section for a definition.

[99] Linuxquestions.org, "Tar," available at: http://wiki.linuxquestions.org/wiki/Tar.

[100] Bibliothèque nationale de France, "The WARC File Format (ISO 28500) - Information, Maintenance, Drafts," available at: http://bibnum.bnf.fr/WARC/.

[101] Descriptions of the BagIt specification are available at: http://en.wikipedia.org/wiki/BagIt; and http://tools.ietf.org/html/draft-kunze-bagit-10.

- *Dissemination Information Packages (DIPs)*

These models build upon each other and for the purposes of these *Guidelines*, they are discussed as being highly interrelated. See the *References* below for more information about each of these package concepts and models.

The initial packaging of digital newspapers for long-term preservation (i.e., forming a *SIP*) should take into account 1) the information necessary to ultimately produce an *AIP*, and 2) any and all information that is essential to restoring the institution's digital news content as *DIP*s in the event of later loss or corruption (part of the goal achieved through *PDI*).

The work involved in packaging content for preservation can vary widely across a spectrum of activity. At the highest end, institutions may specifically follow such packaging standards as those put forward in *OAIS*, using unique identifiers with various metadata or management tools, and packaging content in lossless packaging formats to ensure long-term stability. These *OAIS* concepts and information models can be a "high-bar" standard to achieve. If an institution has the resources, creating such packages for digital newspaper collections will serve the institution well in the long-term.

Curators at smaller and/or less resourced institutions can package digital news content effectively using practices that do not require comprehensive understanding of all the relevant standards. If an institution has followed the lightweight practices recommended in previous sections (particularly inventorying digital news content, organizing, and packaging metadata) that institution will be ready to create sufficient "preservation ready" packages to fulfill the goal of long-term preservation. More on this under *Essential Readiness* below.

For well-resourced institutions, packaging digital newspapers for local preservation may include designating and applying a *unique identifier (UID)* scheme and using a *Name Assigning Authority* for digital newspaper content. *UID*s (which can differ from persistent URLs and do not always rely directly on filenames) are algorithmically assigned identifiers that are unique to the items for which they are created (in this case digital newspaper objects). Once stored and indexed in relation to the objects, they can help curators to *locate, access,* and *manage* digital newspaper files in archival storage environments. In addition, *UID*s can be leveraged via the use of metadata schemas such as *METS*[102] and/or *PREMIS*[103] (see *Section 4. Metadata Packaging for Digital Preservation*) in conjunction with additional scripting to machine-automate many preservation management functions. Much depends on the underlying repository software systems or configurations and how they are designed to facilitate integrations for *UID*s. It is not within the scope of this document to address any particular underlying systems or approaches. Curators should consult their repository system documentation or their repository architects to determine support for *UID*s. Finally, when an institution is packaging digital newspapers to exchange with an external preservation service

[102] Library of Congress, "Metadata Encoding Transmission Standard (METS)," available at: http://www.loc.gov/standards/mets/.

[103] Library of Congress, "Preservation Metadata: Implementation Strategies (PREMIS)," available at: http://www.loc.gov/standards/premis/.

Reference: OAIS Packaging Concepts & Models

*Preservation Description Information (PDI)***:** The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information.

Submission Information Package (SIP): An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Dissemination Information Package (DIP): An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.

provider, it is important to export and include any such existing *UID*s. Note that if these identifiers have been central to a repository infrastructure, they are foundational elements if/when an institution needs to rebuild and restore collections in the future from its preservation copies.

Thinking beyond *UID*s, preservation curators (both local and external) will need ways to validate the integrity (completeness and correctness) of the objects that they receive, monitor, and manage. To that end, an authoritative record or list of all the files and their checksums should be included in the archival package, (see *Section 1. Inventorying Digital Newspapers* and *Section 5. Checksum Management for Digital Newspapers*). *BagIt* and its creation of per-file inventories (which include file extensions) and checksums can be one lightweight approach.

Finally, placing digital newspaper objects or collection units into lossless archival packaging formats helps to maintain their integrity over extended periods of active preservation management and storage media changes. This can provide multiple stakeholders with manageable units of data that can be traced and validated both across and between institutions.

Tools

There are a number of tools and resources that can assist an institution with packaging its digital newspapers for long-term preservation. Selecting the appropriate tools will depend upon the level of resource investment an institution can make - from creating simple, well-documented *TAR* packages to deploying *UID*s (and making use of related *Name Assigning Authorities*) and producing more complex packages (whether *TAR*, *WARC*, or *BagIt*).

Examples of *UID* and packaging tools are provided below with brief descriptions. Usage of these tools is discussed below in the *Essential* and *Optimal* sections.

Unique Identifier (UID) resources that are popular in digital libraries and archives include:

- *Handle System*[104]
- *ARK*[105]
- *NOID*[106]

The *Handle System*, for the purposes of creating local *UID*s, provides local server side application support for maintaining a *Name Assigning Authority (NAA)*,[107] and creating *Name Assigning Authority Numbers (NAAN)* for institutions that can be used as prefixes for unique identifiers. Identifiers in the *Handle* system can be any printable characters in UTF-8 encoding from most major languages written today. *ARK*s, or *Archival Resource Keys*, are unique identifiers that can be created and managed by tools like *NOID*, or *Nice Opaque Identifiers*.

In order to use these tools, an institution should first register with a federated *Name Assigning Authority* like that maintained by the California Digital Library (CDL). Doing so will provide the institution with a NAAN prefix that it can use in conjunction with an *ARK* created by *NOID* to form a *UID* for a collection or set of digital newspaper objects. This *UID* can then be stored with metadata and can be used to manage the collection in archival storage. *METS* and/or *PREMIS* can be especially helpful metadata tools for institutions that have moved to the stage of creating *UID*s for their digital newspapers (more on this below).

Lossless archival packaging formats include:

- *TAR*[108]
- *WARC*[109]
- *METS*[110]
- *PREMIS*[111]
- *BagIt*[112]

TAR, which stands for *Tape ARchive*, is a lossless packaging format that works especially well for encapsulating folders of files and maintaining file system metadata about the objects and structures contained therein. There are a number of standard and open source utilities available for producing and unpacking *TAR* packages. *TAR* packages can be subsequently compressed, but compression is not advised for long-term storage and preservation.

[104] Corporation for National Research Initiatives, "Handle System," available at: http://www.handle.net/.

[105] California Digital Library, "ARK (Archival Resource Key) Identifiers," available at: https://wiki.ucop.edu/display/Curation/ARK.

[106] California Digital Library, "NOID: Nice Opaque Identifier (Minter and Name Resolver)" available at: https://wiki.ucop.edu/display/Curation/NOID.

[107] For a list of current Name Assigning Authority Numbers see, California Digital Library, "Registered Name Assigning Authority Numbers," available at: http://www.cdlib.org/services/uc3/naan_table.html.

[108] Linuxquestions.org, "Tar," available at: http://wiki.linuxquestions.org/wiki/Tar.

[109] Bibliothèque nationale de France, "The WARC File Format (ISO 28500) - Information, Maintenance, Drafts," available at: http://bibnum.bnf.fr/WARC/.

[110] Library of Congress, "Metadata Encoding Transmission Standard (METS)," available at: http://www.loc.gov/standards/mets/.

[111] Library of Congress, "Preservation Metadata: Implementation Strategies (PREMIS)," available at: http://www.loc.gov/standards/premis/.

WARC, which stands for *Web ARchive*, is a lossless packaging format typically used for encapsulating harvested websites. It improves upon an earlier *ARC* format by better describing the web resources it contains and enabling improved harvesting, sharing and access. Standard web-crawling software such as Heritrix can produce *WARC* packages.

METS and PREMIS were previously described in the *Section 4. Metadata Packaging for Digital Newspapers*. These standards can be used to maintain linkages between associated digital newspaper objects, their metadata, and any assigned *UID*s. They can also help to record preservation actions taken on archived objects.

Finally, *BagIt* is a packaging specification that can be applied to a digital newspaper collection at any level in a collection hierarchy to produce an inventory and checksums for that content. *BagIt* can be especially helpful for auditing locally preserved content or verifying the integrity of collection content when exchanged with an external preservation service provider. As mentioned in *Section 5. Checksum Management for Digital Newspapers*, there are GUI-based tools like *Bagger*[113] that can make the implementation of *BagIt* relatively painless.

Readiness Spectrum

Ultimately, what an institution packages is what that institution will get back out of its archival storage for restoration purposes. The packaged content must be adequate to restore objects and collections in the event of corruption or loss. All of the previous sections in these *Guidelines* have built toward this step of packaging with explicit attention to recovery/reconstitution of collections.

For smaller or under-resourced institutions, this packaging phase may be as simple as including written instructions for three scenarios: 1) how to retrieve a single file or subset of files to restore damaged local copies; 2) how a collection can be restored within or re-imported into its repository system/environment; and 3) how the institution should go about reproducing derivative access copies of its collection content from their master files (if those differ). Applying simple and reliable archival packaging formats such as *BagIt* and *TAR* (uncompressed) could then prove helpful for on-going management purposes and assurances.

For institutions with greater resources, efforts should be taken to achieve optimal *Preservation Description Information (PDI)* and to enrich *Submission Information Packages (SIPs)* with as much information as possible to achieve a sufficient *Archival Information Package (AIP)*. If long-term preservation is being pursued locally, the use of unique identifiers (*UIDs*) should be considered, ideally in combination with preservation metadata schemas such as *METS* and/or *PREMIS*. Likewise, deploying reliable archival packaging formats such as *BagIt* and/or *TAR*, and more complex formats such as *WARC* (which is particularly useful for website-based born-digital newspapers), could prove helpful to future restoration activities.

For institutions both large and small that are partnering with external preservation providers, *BagIt* can facilitate the trustworthy transfer of any such packaged digital newspaper

[112] Descriptions of the BagIt specification are available at: http://en.wikipedia.org/wiki/BagIt and http://tools.ietf.org/html/draft-kunze-bagit-10.

[113] Library of Congress, "Bagger," available at: http://libraryofcongress.github.io/bagger/.

collections. *BagIt* ideally should be applied prior to any additional packaging using *TAR* or *WARC*. Use of *BagIt* prior to packaging with *TAR* is logical because it enables inventorying of the subsequent *TAR* contents on a per-file basis rather than inventorying of the one *TAR* file itself. Use of *BagIt* before or after any packaging with *WARC* is as yet largely unexplored, but follows from the same logic.

Essential Readiness

As mentioned in *Section 2. Organizing Digital Newspapers for Preservation* as well as above in this Section, institutions with less resources to put behind packaging their digital newspapers according to high-bar standards should aim to document collection structures and how these can and should be re-constituted in the event of loss or corruption. In addition to the documentation efforts recommended in *Section 1. Inventorying Digital Newspapers for Preservation*, some additional factors to be considered in the creation of such documentation include:

- Document *where* and *on what media* the master preservation copies for any given collection currently reside at the local institution;
- Document any preferred form of media or means for receiving back preservation copies for the sake of a recovery and restoration (SFTP,[114] hard drive, CD, DVD, etc.);
- Document instructions on how to work with the above-mentioned media or mechanism to retrieve preservation copies from a local preservation system or an external preservation service provider;
- Document step-by-step instructions for how to restore collections from their preservation copies (including how to make use of any previously developed inventories and/or assigned unique identifiers); and
- Document guidance on how to produce derivative access copies from restored preservation copies (in case access copies are lost or are not preserved)

This documentation should be produced and saved in an open and non-proprietary file format (e.g., *Open Document Format*,[115] *PDF*,[116] or even a simple *TXT* file). It can and should be stored in its own clearly labeled folder or alongside the collection folders as a README with its own clearly labeled filename.

Once an institution has documented its digital newspaper collections for packaging purposes, the institution can consider encapsulating its digital newspaper collections in a lossless archival packaging format.

The easiest and most reliable packaging format for less-resourced institutions to begin working with is *TAR*.[117] Packaging digital newspaper collections as *TAR* files can facilitate the management of digital newspaper collections as

[114] A description of the SSH File Transfer Protocol (SFTP) and its differences with both SSH and FTP are available at: http://en.wikipedia.org/wiki/SSH_File_Transfer_Protocol.

[115] OASIS, "OpenDocument Format," available at: http://www.opendocumentformat.org/.

[116] AIIM Standards Wiki, "PDF/A," available at: http://pdf.editme.com/PDFA.

[117] There are several platform-specific third-party tools to create and unarchive TAR packages, all of which should be investigated over time for their level of upkeep and stability in correctly handling the format.

units of related data within an archival storage environment and across migrations of storage media over time. Institutions creating a *TAR* package for digital newspapers should consider doing so at whatever level of aggregate is most representative of their intellectual organizations and structures - be that at a title, volume, or issue level. They should also clearly label the TAR file without overwriting its extension. Consider some of the file-naming conventions provided in *Section 2. Organizing Digital Newspapers for Preservation*.

Optimal Readiness

Institutions with more resources can take guidance from OAIS concepts such as *Preservation Description Information (PDI)* as they form *Submission Information Packages (SIPs)* and *Archival Information Packages (AIPs)*. These packages should include information such as:

- *Reference Information* that can assist the retrieval of the collection from within an archival storage environment, be that local or an external preservation service provider (unique identifiers can be helpful here);
- *Context Information* that can help to explain the relationship between a digital newspaper collection and its environment and other digital newspaper collections (METS can be helpful for defining these relationships);
- *Provenance Information* that can help archival managers and users understand the chain of stewardship for the collection and what sort of preservation actions have taken place over time (*METS* and *PREMIS* are increasingly being used for recording such information);
- *Fixity Information* that can assist with auditing a collection and its digital objects over the course of its archival management (*BagIt* is one simple approach to creating and storing per-file fixity with a collection); and
- *Access Rights* that can help curators or external preservation service providers understand what levels of access and handling are permitted for the collection as it is managed over time (generic descriptive metadata can make this evident and can also be used within *METS*).

As mentioned in *Section 4. Metadata Packaging for Digital Newspapers*, institutions with greater resources should consider implementing *METS* and/or *PREMIS* for their digital newspaper content because these standards explicitly retain linkages between associated digital newspaper objects and their metadata. *METS* and *PREMIS* can also accommodate and leverage unique identifiers (*UID*s) and operate as *XML* schemas that help with automating various archival management and access processes for any digital objects that they describe and encapsulate. Making use of *METS* and/or *PREMIS* in specification-conformant ways is an ideal way to package digital newspaper collections.

Unique identifiers (*UID*s) should be a priority for institutions using *METS* and/or *PREMIS*. *UID*s require a *Name Assigning Authority (NAA)* that can register the institution and mint a unique identifier. As mentioned above, one of the increasingly popular *NAA*s is the one maintained by the California Digital Library (CDL) and mirrored at the National Library of Medicine and the Bibliothèque Nationale de France. Each institution that registers is given a *Name Assigning Authority Number (NAAN)* that it can use as a prefix for the identifier. As mentioned earlier, *Archival Resource Keys (ARKs)* or *Handles* can be used to produce *UID*s for digital newspaper collections or individual digital newspaper objects. *NOID*, or *Nice Opaque Identifier*, is a service that can assist with minting unique *ARK*s or *Handles* that can then be coupled with the institution's *NAAN*.

Once an institution has assigned identifiers and associated metadata with its newspapers, it can encapsulate news data by applying *BagIt* and placing the collections in an archival format such as *TAR* or *WARC*. *BagIt* produces a full inventory of included files and per-file checksums that assists in validating content during exchanges and on-going audits. The best packaging model may differ across born-digital or digitized content types. For example, *TAR* works very well for packaging well-organized and hierarchical folders of digitized newspaper collections. *WARC* on the other hand is geared primarily toward packaging website oriented collections.

Section 7. Additional Considerations

7.1. Creation & Acquisition

As mentioned in the *Introduction*, the *Guidelines* are designed to address digital newspaper preservation and curation - not creation and acquisition. The issues involved in creating and acquiring digital newspapers warrant their own "Guidelines" documentation (and indeed, there are many standards and guidelines available on these topics: see e.g., *National Digital Newspaper Program Technical Guidelines*[118] for digitized newspapers, the *Federal Agencies Digitization Guidelines*,[119] and the *International Internet Preservation Consortium Web Archiving*[120] resources that lend themselves to born-digital newspapers, to name just a few) - however, there are components of digital newspaper preservation and curation that should be addressed as an institution initiates digital newspaper collections of any type. This *Additional Considerations* section, first and foremost, aims to provide an overview of the creation and acquisitions concerns an institution should consider as part of its overall approach to digital news preservation.

There are a number of preservation-oriented practices that stewards of digital news collections can begin to integrate and incorporate at the creation and acquisition phase. For example:

- Curators can begin to record a full set of inventory data (see *Section 1. Inventorying Digital Newspapers for Preservation*) as digital newspaper content arrives via hard drives or FTP retrievals (or other mechanisms) from vendors, institutional digitization units, or publishers and other data providers;
- Curators and their technical staff members can coordinate to analyze and record file format information at creation/acquisition through whatever form(s) best suits the institutional framework and its new collections: e.g., lightweight manual processes, automated workflows using open source tools, or some combination of both (see *Section 3. Format Management for Digital Newspapers*);
- Curators and metadata specialists can negotiate with digitization vendors, institutional digitization units, and data providers (including publishers) to provide the institution with the most complete, usable and conformant set of metadata elements and formats possible. This can help to ensure integration with supported schemas and metadata workflows (see *Section 4. Metadata Packaging for Digital Newspapers*). The *National Digital Newspaper Program's Technical*

[118] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants," August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

[119] Federal Agencies Digitization Guidelines Iniative, "Digitization Guidelines," available at: http://www.digitizationguidelines.gov/.

[120] IIPC, "Web Archiving," available at: http://netpreserve.org/web-archiving/overview.

Guidelines[121] illustrate how specific metadata requirements were organized and made available for reference and implementation by digitization service providers;

- Curators can request checksum information from digitization vendors, institutional digitization units, and any other data providers prior to their hand-off of news collections. These checksums can then be used to verify that the initial transfer, as well as any subsequent work performed on the content, has not altered the original objects (see *Section 5. Checksum Management for Digital Newspapers*);
- Curators can proactively request that digitization vendors, institutional digitization units, or publishers/born-digital news producers create and store newspaper content in established structures that conform to the institution's preferred file and folder-naming conventions (see *Section 2. Organization for Digital Newspapers*); and
- Finally, curators can request that those producing or providing the institution with born-digital newspaper content provide the explicit permissions needed by the institution in order to preserve the content over time (see *Section 6. Packaging Digital Newspapers for Preservation* and the section below on *Preservation Partners and Permissions*).

7.2. Preservation Partners & Permissions

The intellectual property and copyright issues inherent to news content can present significant preservation challenges, particularly when permissions are not carefully negotiated and recorded as content is created or acquired. Understanding the implications of reproducing and/or providing access to newspaper content requires deeper legal knowledge than can be conveyed in these brief *Guidelines*.

Most institutions working with digital news content have collections (digitized and/or born-digital) that are not in the public domain and for which there may be explicit access and/or reproduction restrictions. Institutions should seek out expert legal advice regarding any risks that might be involved in replicating digital news content for preservation purposes and/or broadening the chain of custody for these objects by working with external preservation vendors or preservation networks.

With (and sometimes without) full clearance for copyright permissions and the proper partner agreements in place, institutions can and do partner with one another and with external service providers to preserve copies of their digital newspapers. The current *U.S. Copyright Act*[122] does not have explicit provisions in place to guide the use of reproduction as a preservation strategy for published digital works. Currently, it permits the creation and management of three copies under specific conditions for unpublished works. The *U.S. Copyright Act* was implemented in an analog era, and updating portions of it (especially Section 108) to address digital content has already been the subject of much debate. In 2008, the *Section 108 Study Group*[123] made a series of recommendations for reforming

[121] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants," August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

[122] US Government, "Copyright Law of the United States," available at: http://www.copyright.gov/title17/.

[123] US Government, "Section 108 Study Group," available at: http://www.section108.gov/.

the current copyright law that would grant libraries and archives the legal rights to make multiple copies of digital content for preservation purposes. In April 2013, the House Judiciary Committee announced plans to hold a series of comprehensive hearings to determine how the current *U.S. Copyright Act* should be amended for the digital age. As of January 2014, these hearings have begun, focusing on the *Fair Use Doctrine*.[124]

7.3. Distribution vs. Back Up

Differentiating between creating back-ups and engaging in preservation is crucial. Backup systems are for the most part non-intelligent and will merely produce a limited set of direct copies of assigned data - regardless of its sustainability or integrity. Many backup systems have the built-in tendency for overwriting older healthy copies of data with more recent (including corroded) bits, particularly if data is not being monitored proactively by a curator. Backup systems may instill a false sense of protection and security (i.e., let the machines handle it). Most backup systems do not keep multiple copies of data in sync with one another in any audited sense. Backup systems often are maintained in close proximity to master copies *and* by the same staff members; making it plausible that one catastrophic event or human act of malice or error could destroy or corrupt all copies.

Digital news curation - like other digital content curation - demands attention to content risks. Sound practice recommends providing for at least three copies of data that do not share a similar set of natural or man-made threats. There are *distributed digital preservation (DDP) systems* that perform replication across a limited geographic distance - for example one institution, with replications at multiple sites within the same city or neighboring locale or region (e.g., University of North Texas and their Coda Repository); *DDP systems* that support three replications of data across a large nation (e.g., Chronopolis); as well as *DDP systems* that provide for up to seven replications of data across multiple continents (e.g. MetaArchive Cooperative). These are just a few examples, and ones that have explicitly benchmarked their systems for preserving digital newspaper data as part of the Chronicles in Preservation Project. It is worth noting that each of these three systems have also worked together to replicate each others' digital collections to demonstrate the importance of replicating content across multiple heterogeneous storage infrastructures.

7.4. Change Management

Even the most carefully curated collections experience change for a variety of reasons, and those changes can impact the preservation outlook for the collection. For example, after a preservation package has been created, ingested into a preservation repository, and stored, a curator may need to add or change a file to that package, particularly if it is a meaningful body of content (i.e., a collection, not just a batch of unrelated files). Perhaps a file will need to be re-digitized and re-incorporated into the preserved collection or a metadata record or *OCR* file will need correction. Such changes can be significant and may impact the preservation package.

[124] House Judiciary Committee, "The Scope of Fair Use," January 2014, available at: http://judiciary.house.gov/index.cfm/2014/1/the-scope-of-fair-use.

Prior to moving content into preservation storage, institutions should grapple with the implications of file changes and how to document and store changed files within a preservation repository and its given workflows. In some cases this may mean that the file name is ascribed (manually or by the system) with a new creation date or an incremented version count. In other cases there may be simple non-numerical renaming conventions that can be ascribed to indicate the change. Relationships (inter-linkages), and/or inheritances of any previously assigned metadata also need to be considered.

As a first step, institutions must decide whether updated files will simply replace any previously collected and stored file. If so, a new file may not require the ascription of any change information. However, the issue of information loss and irretrievability in such over-writing should be weighed carefully. Growth of collections is also important when it comes to change management - particularly when working with external preservation partners. Such on-going exchanges of new and/or refreshed data needs to be well coordinated, synchronized, and documented across all the partners (metadata can sometimes help with this but other documents may also play a role). Whatever the situation, a reliable, consistent, and clearly communicated policy and practice should be the goal. Creation or modified-since date information (YYYY_MM_DD or YYYYMMDD) is one of the easiest solutions and has already been particularly useful (and fairly standard) when it comes to distinguishing digital newspaper content - it also has some use value for indicating changes (as already mentioned above).

7.5. Preservation Monitoring

Once content has been packaged for long-term preservation (see *Section 6. Packaging Digital Newspapers for Preservation*) and moved into archival storage/workflows, there are a number of package elements that need to be monitored at key intervals. For example, an audit schedule should be established to ensure regular per-file fixity and to make fixity comparisons across any replicated copies of files (see the section above on *Distribution vs. Back Up*). Institutions must determine what role they will play in this ongoing monitoring of content, including whether they are willing to outsource these tasks to external groups (e.g., preservation service providers) and what reports/documentation they expect to review. There are a number of questions to consider, including the following:

- What is the supported policy/practice for repairing a corrupted file and is this handled in a traceable and authorized manner?
- Who has permissions to access, manage and if necessary update preservation copies of digital newspaper data in approved ways?
- What types of network analysis, security measures, and system redundancies are in place to guard against disasters, unwarranted intrusions, and general accidents?
- How are incidents logged and reported?

Institutions that do not have direct control over their preservation storage environments should make sure that service providers that they are partnering with provide an appropriate degree of reporting and/or assurances around these and other sorts of monitoring concerns.

The *Reference Model for an Open Archival Information System (OAIS)*[125] and the *ISO 16363:2012 Audit and certification of trustworthy digital repositories*[126] set forth a broad range of monitoring factors and approaches that should be in place and accounted for. However, as stressed throughout these *Guidelines*, these standards and metrics can be followed and implemented in measured and incremental fashions. At this admittedly early stage in the formation of the digital preservation field, all efforts to preserve digital newspapers (and any other digital content) should be pursued with the understanding that digital preservation is quickly evolving. More than any full-blown implementation of standards, the most important task for memory institutions is to stay vigilant and avoid falling prey to the "out-of-sight, out-of-mind" mentality that can so easily come with this digital terrain. Take responsibility, maintain control, ask questions always, and demand information if need be.

7.6. Recovering Digital Newspapers from Preservation

The *Guidelines* have emphasized, particularly in *Section 2. Organizing Digital Newspapers for Preservation* and *Section 6. Packaging Digital Newspapers for Preservation*) that the time and work associated with recovery depends largely upon the work an institution does to logically structure, document and package its digital newspaper collections. Loss or corruption scenarios may involve one or a small number of files or they may involve whole collections. The institution as a data provider needs to know how and where to turn to issue a request for any preserved copies of data - be that in coordination with local archival management or an external preservation service provider. There may be specific request channels and protocols that need to be observed and followed. The institutional owner of the data should have all the necessary identifying information as it corresponds to the data stored so that proper identifications and timely retrievals can take place.

Though the *Reference Model for an Open Archival Information System (OAIS)* primarily refers to this stage of activity in terms of Access and *Dissemination Information Packages (DIPs)* as they relate to end-user access requests, the information can be helpful for understanding requests for data from an archive as a bounded and segregated activity that should occur using resources not necessarily shared by those reserved for archival management. In other words there should be server/storage resources assigned for copying *AIP*s or objects within *AIP*s to in order to facilitate a retrieval of this data for recovery purposes - this might be an FTP enabled server, an external hard drive, or some other portable media for shipping/delivery purposes. Negotiating this media and delivery mechanism is important prior to facing an actual recovery scenario because it can determine what sorts of support needs to be in place at the receiving institution.

The institutional owner of this digital newspaper content should include within a *DIP* everything it needs to verify the correctness of the file(s) and their bit integrity. These *Guidelines* have consistently aimed to

[125] CCSDS, "Reference Model for an Open Archival Information System (OAIS) – Magenta Book," available at: http://public.ccsds.org/publications/archive/650x0m2.pdf.

[126] CCSDS, "ISO 16363:2012 Audit and certification of trustworthy digital repositories – Magenta Book," available at: http://public.ccsds.org/publications/archive/652x0m1.pdf.

facilitate this objective by advocating for various documentary and metadata approaches that can disclose filenames and their linkages to metadata and checksum information. In the event of a full collection recovery, if metadata has been properly stored along with both the collection files and any *unique identifiers (UIDs)* that the local repository system uses to reintegrate collection content, this set of activities should be less cumbersome. The owning institution should ideally test such recoveries as part of its initial archival storage efforts to ensure a seamless integration between archival management and recovery.

References

Below are the tools and resources referenced in the *Guidelines*.

Active Paper - Olive Software, “Active Paper Archive,” available at: http://www.olivesoftware.com/activepaper-archive.html.

Adobe Image Processor - Adobe Systems Incorporated, “Image Processor Script,” available at: http://tv.adobe.com/watch/understanding-adobe-photoshop-cs6/image-processor-script/. Adobe’s software requires a paid license, but it is widely used for bulk image format migrations.

ALCTS - Association for Library Collections & Technical Services, “Newspaper IG,” available at: http://www.ala.org/alcts/mgrps/ig/ats-dgnews.

ArchivalWare - “Libraries,” available at: http://www.archivalware.net/libraries.

ARK - California Digital Library, “ARK (Archival Resource Key) Identifiers,” available at: https://wiki.ucop.edu/display/Curation/ARK.

Automator - Apple, “Mac Basics: Automator,” available at: http://support.apple.com/kb/ht2488.

Bagger - Library of Congress, “Bagger,” available at: http://libraryofcongress.github.io/bagger/.

BagIt - Descriptions of the BagIt specification are available at: http://en.wikipedia.org/wiki/BagIt and http://tools.ietf.org/html/draft-kunze-bagit-10.

bagit-python - Library of Congress, “bagit-python,” available at: http://libraryofcongress.github.io/bagit-python/.

Bulk Rename Utility – “Bulk Rename Utility Homepage,” available at: http://www.bulkrenameutility.co.uk/Main_Intro.php.

Chronicles in Preservation - MetaArchive Cooperative, “Chronicles in Preservation,” available at: http://metaarchive.org/neh/index.php/Main_Page .

code4lib - Code4Lib, “Homepage,” available at: http://www.lsoft.com/scripts/wl.exe?SL1=CODE4LIB&H=LISTSERV.ND.EDU.

cron - Linuxquestions.org, “Scheduling tasks,” available at: http://wiki.linuxquestions.org/wiki/Scheduling_tasks.

CSS - W3C, “Cascading Style Sheets,” available at: http://www.w3.org/Style/CSS/.

CSV - Wikipedia, “Comma-separate values,” available at: http://en.wikipedia.org/wiki/Comma-separated_values.

digipres - American Library Association, "digipres- Digital Preservation," available at: http://lists.ala.org/sympa/info/digipres.

digital-curation - Google Groups, "Digital Curation-Google Groups," available at: https://groups.google.com/forum/#!forum/digital-curation.

DROID - UK National Archives, "Droid," available at: http://digital-preservation.github.io/droid/.

DSpace - "DSpace Homepage," available at: http://www.dspace.org/.

Dublin Core - Dublin Core Metadata Initiative, "DCMI Home," available at: http://dublincore.org/.

Exiftool - Phil Harvey, "Exiftool," available at: http://www.sno.phy.queensu.ca/~phil/exiftool/.

FADGI - Federal Agencies Digitization Guidelines Initiative, "Digitization Guidelines," available at: http://www.digitizationguidelines.gov/.

Fair Use Doctrine - House Judiciary Committee, "The Scope of Fair Use," January 2014, available at: http://judiciary.house.gov/index.cfm/2014/1/the-scope-of-fair-use.

FCLA Description Service - Florida Center for Library Automation, "Description Service," available at: http://description.fcla.edu/.

Fedora - "Fedora Repository," available at: http://fedora-commons.org/.

find - LinuxQuestions.org, "find," available at: http://wiki.linuxquestions.org/wiki/Find.

Finder - Apple, "Mac OS X: Finder Basics," available at: http://support.apple.com/kb/VI209.

file - LinuxQuestions.org, "file," available at: http://wiki.linuxquestions.org/wiki/File_%28command.

File Explorer - Microsoft, "How to work with files and folders," available at: http://windows.microsoft.com/en-us/windows-8/files-folders-windows-explorer.

File Format Wiki - Archiveteam, "Archiveteam File Format Wiki," available at: http://fileformats.archiveteam.org/.

File naming - Recommendations from the State Archives of North Carolina, "Digital Records Policies & Guidelines: Filenaming," available at: http://www.ncdcr.gov/archives/ForGovernment/DigitalRecords/DigitalRecordsPoliciesandGuidelines.aspx#filenaming.

FITS - Harvard University Library Office for Information Systems, "File Information Tool Set," available at: http://project.iq.harvard.edu/fits.

Fixity - AVPreserve, "Fixity," available at: https://github.com/avpreserve/fixity.

GPRename - "GPRename," available at: http://gprename.sourceforge.net/.

Handles - Corporation for National Research Initiatives, “Handle System,” available at: http://www.handle.net/.

hashdeep - Jesse Kornblum, “md5deep and hashdeep,” available at: http://md5deep.sourceforge.net/.

Heritrix - Internet Archive, “Heretrix,” available at: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix.

HTML - W3C, “HTML,” available at: http://www.w3.org/html/.

IFLA - International Federation of Library Associations, “Newspapers Section,” available at: http://www.ifla.org/newspapers.

IIPC - IIPC, “Web Archiving,” available at: http://netpreserve.org/web-archiving/overview.

ImageMagick - ImageMagick Studio LLC, “ImageMagick,” available at: http://www.imagemagick.org/.

ISO 8601 - A description of the rationale and usage of the international date standard is available at: http://en.wikipedia.org/wiki/ISO_8601.

ISO 14721 - Consultative Committee for Space Data Systems, CCSDS 650.0-M-2: Reference Model for an Open Archival Information System (OAIS): Magenta Book, June 2012, available at: http://public.ccsds.org/publications/archive/650x0m2.pdf.

ISO 16363 - International Organization for Standardization, ISO 16363:2012: Space data and information transfer systems -- Audit and certification of trustworthy digital repositories, February 2012, available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

JavaScript - Mozilla Developer Network, “JavaScript,” available at: https://developer.mozilla.org/en-US/docs/Web/JavaScript.

JHOVE2 - “JHOVE2,” available at: https://bitbucket.org/jhove2/main/wiki/Home.

JPEG2000 - Joint Photographic Experts Group, “JPEG2000,” available at: http://www.jpeg.org/jpeg2000/index.html.

Levels of Preservation - Library of Congress, “NDSA Levels of Preservation,” available at: http://www.digitalpreservation.gov/ndsa/activities/levels.html.

locate - LinuxQuestions.org, “locate,” available at: http://wiki.linuxquestions.org/wiki/Locate.

ls - LinuxQuestions.org, “ls,” available at: http://wiki.linuxquestions.org/wiki/Ls.

Md5 - A description of the md5 algorithm is available at: http://en.wikipedia.org/wiki/md5.

md5sum - Linuxquestions.org, “Md5sum,” available at: http://wiki.linuxquestions.org/wiki/Md5sum.

Metadata Extraction Tool - National Library of New Zealand, “Metadata Extraction Tool,” available at: http://meta-extractor.sourceforge.net/.

METS - Library of Congress, “Metadata Encoding Transmission Standard (METS),” available at: http://www.loc.gov/standards/mets/.

METS-ALTO - Library of Congress, “National Digital Newspaper Program: Digitizing Microfilm and Optical Character Recognition,” December 2012, available at: http://www.loc.gov/ndnp/guidelines/digitizing.html.

MIX - Library of Congress, “Metadata for Images in XML Standard,” available at: http://www.loc.gov/standards/mix/.

MODS - Library of Congress, “Metadata Object Description Schema,” available at: http://www.loc.gov/standards/mods/.*mv* - LinuxQuestions.org, “Mv,” available at: http://wiki.linuxquestions.org/wiki/Mv.

Name Assigning Authority Numbers - For a list of current Name Assigning Authority Numbers see, California Digital Library, “Registered Name Assigning Authority Numbers,” available at: http://www.cdlib.org/services/uc3/naan_table.html.

Newslib - “Homepage,” available at: http://www.ibiblio.org/slanews/NewsLib/newsliblyris.html.

Nautilus - The GNOME Project, “Nautilus,” available at: https://wiki.gnome.org/action/show/Apps/Nautilus.

NDNP Technical Guidelines - Library of Congress, “The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants,” August 2012, available at: http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf.

NDSA - Library of Congress, “National Digital Stewardship Alliance Homepage,” available at: http://www.digitalpreservation.gov/ndsa/.

NOID - California Digital Library, “NOID: Nice Opaque Identifier (Minter and Name Resolver)” available at: https://wiki.ucop.edu/display/Curation/NOID.

ODF - OASIS, “OpenDocument Format,” available at: http://www.opendocumentformat.org/.

PDF - AIIM Standards Wiki, “PDF,” available at: http://pdf.editme.com/PDFREF.

PDF/A - AIIM Standards Wiki, “PDF/A,” available at: http://pdf.editme.com/PDFA.

PREMIS - Library of Congress, “Preservation Metadata: Implementation Strategies (PREMIS),” available at: http://www.loc.gov/standards/premis/.

PRONOM - UK National Archives, “PRONOM Technical Registry,” available at: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

SAA - Society of American Archivists, “Homepage,” available at: http://www2.archivists.org/.

Section 108 Study Group - US Government, “Section 108 Study Group,” available at: http://www.section108.gov/.

sed - LinuxQuestions.org, “Sed,” available at: http://wiki.linuxquestions.org/wiki/Sed.

SFTP - A description of the SSH File Transfer Protocol (SFTP) and its differences with both SSH and FTP are available at: http://en.wikipedia.org/wiki/SSH_File_Transfer_Protocol.

sha-1 - A description of the sha-1 algorithm is available at: http://en.wikipedia.org/wiki/sha1.

Sha-256 - A description of the sha-2 algorithm, of which sha-256 is a part, is available at: http://en.wikipedia.org/wiki/sha2.

sha1sum - Linuxquestions.org, “Sha1sum,” available at: http://wiki.linuxquestions.org/wiki/Sha1sum.

Shell Scripts - LinuxCommand.org, “Writing Shell Scripts,” available at: http://linuxcommand.org/lc3_writing_shell_scripts.php.

Sustainability of Formats - Library of Congress, “Library of Congress Sustainability of Formats,” available at: http://www.digitalpreservation.gov/formats/.

tar - Linuxquestions.org, “Tar,” available at: http://wiki.linuxquestions.org/wiki/Tar.

Task Scheduler - Microsoft, “Schedule a task,” available at: http://windows.microsoft.com/en-US/windows/schedule-task.

TIFF - Adobe Systems Incorporated, “TIFF,” available at: http://partners.adobe.com/public/developer/tiff/index.html.

Tika - Apache Software Foundation, “Apache Tika,” available at: https://tika.apache.org/. Apache Tika has a number of use cases – see here: http://www.openplanetsfoundation.org/blogs/2013-05-20-apache-tika-file-mime-type-identification-and-importance-metadata.

TRAC - Center for Research Libraries, “Trustworthy Repositories Audit & Certification: Criteria & Checklist,” February 2007, available at: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

UDFR - California Digital Library, “Unified Digital Format Registry (UDFR),” available at: http://udfr.cdlib.org/.

US Copyright Act - US Government, “Copyright Law of the United States,” available at: http://www.copyright.gov/title17/.

WARC - Bibliothèque nationale de France, “The WARC File Format (ISO 28500) - Information, Maintenance, Drafts,” available at: http://bibnum.bnf.fr/WARC/.

Xena - National Archives of Australia, “Xena – Digital Preservation Software,” available at: http://xena.sourceforge.net/.

XHTML - W3C, “W3C XHTML2 Working Group Home Page,” available at: http://www.w3.org/MarkUp/.

XML - W3C, “XML 1.0,” available at: http://www.w3.org/TR/REC-xml/.